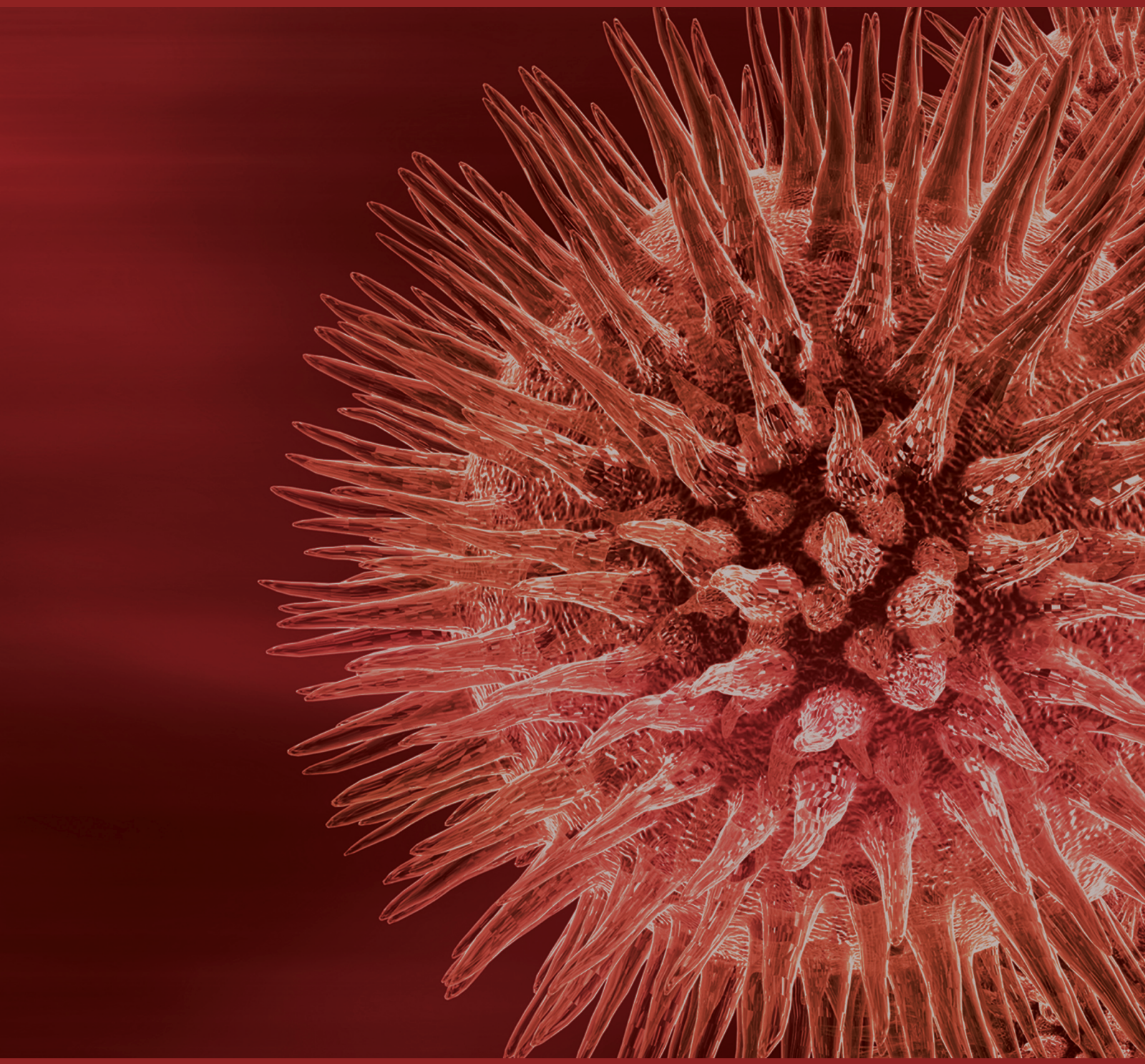
BioMed Research International

Intelligent Informatics in Biomedicine

Guest Editors: Hao-Teng Chang, Raffaele A. Calogero, Sorin Draghici, Oliver Ray, and Tun-Wen Pai



Intelligent Informatics in Biomedicine

BioMed Research International

Intelligent Informatics in Biomedicine

Guest Editors: Hao-Teng Chang, Raffaele A. Calogero, Sorin Draghici, Oliver Ray, and Tun-Wen Pai

Contents

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 185839, 2 pages
http://dx.doi.org/10.1155/2013/185839

Editorial
Intelligent Informatics in Biomedicine

Hao-Teng Chang,[1] Raffaele A. Calogero,[2] Sorin Draghici,[3] Oliver Ray,[4] and Tun-Wen Pai[5,6]

[1] *Graduate Institute of Basic Medical Science, College of Medicine, China Medical University, Taichung, Taiwan*
[2] *Department of Clinical and Biological Sciences, Torino University, Torino, Italy*
[3] *Department of Computer Science, Wayne State University, Detroit, MI 48202, USA*
[4] *Department of Computer Science, University of Bristol, Bristol BS8 1UB, UK*
[5] *Department of Computer Science and Engineering, National Taiwan Ocean University, No. 2 Peining Road, Keelung 20224, Taiwan*
[6] *Center for Marine Bioenvironment and Biotechnology, National Taiwan Ocean University, No. 2 Peining Road, Keelung 20224, Taiwan*

Correspondence should be addressed to Tun-Wen Pai; twp@mail.ntou.edu.tw

Received 27 March 2013; Accepted 27 March 2013

In the postgenomic era, hundreds of thousands of biological datasets including genetic information of various species, gene expression profiles, metabolomes, proteomes, and even molecular imaging are published in public domains. One, maybe the most, important reason to conduct various genome projects is to translate useful relevant information to biomedical research and finally to clinical applications. From bench to bed, bioinformatics researches have presented strong and powerful potential to accelerate the analyses of comprehensive and complicated datasets. To establish a forum for gathering scientists from multidisciplinary fields such as biology, medicine, computer science, statistics, and informatics, Dr. Hui-Huang Hsu, Dr. Tun-Wen Pai, Dr. Oliver Ray, and Dr. Hao-Teng Chang are continuously involved in organizing International Workshop on Intelligent Informatics in Biology and Medicine (IIBM) starting from 2008, which is used to be held in conjugation with International Conference on Complex, Intelligent, and Software Intensive Systems (CISIS).

In these years, IIBM successfully brings together computer scientists, biologists, statisticians, and medical doctors to present and discuss current topics on intelligent informatics in biology and medicine. Although the complexity and volume of experimental data from next generation sequencing and mass spectrometry technologies increase dramatically, many various sophisticated computational techniques have been designed and developed. These methodologies are able to support new detection techniques that are developed to improve the quality of healthcare and medicine. Advances in information technologies, indeed, facilitate and accelerate research from basic to clinical investigations in terms of translational medicine.

To record the ideas of talents and gather more contributions to these fields, this special issue was launched and supported by this journal. This special issue focuses on the challenges and solutions for information process with an emphasis on forthcoming high throughput technologies and biomedicine systems, which will provide opportunities for academics and industrial professionals to discuss the latest issues and progresses in the area of biomedicine. This special issue contains 9 papers which were selected from 24 submissions. These papers address the development and application of data-analytical methods, algorithm development, mathematical modeling, and computational simulation techniques to the biomedicine applications.

In "*Time series expression analyses using RNA-seq: a statistical approach*," S. Oh et al. apply three real datasets and simulation studies to demonstrate the utility of statistical evolutionary trajectory index, autoregressive time-lagged

regression, and hidden Markov model approaches for RNA-seq datasets in temporal version.

In "*Gene entropy-fractal dimension informatics with application to mouse-human translational medicine*," T. Holden et al. compare the Shannon entropy and fractal dimension of some DNA sequences computed from different mammalian species. The obtained values were plotted in a 2D map, and the distance between points on these maps for corresponding mRNA sequences in different species is used to study evolutionary topics.

In "*State-of-the-art fusion-finder algorithms sensitivity and specificity*," M. Carrara et al. utilize seven existing state-of-arts gene fusion detection tools. Their strategy is to simulate some gold-standard data and then compare sensitivity and specificity of the detection tools. The experimental results obtained using synthetic and real datasets suggest that synthetic datasets encompassing fusion events may not fully catch the complexity of RNA-seq experiments, and most fusion detection tools are still limited in sensitivity or specificity.

In "*On the structural context and identification of enzyme catalytic residues*," Y.-T. Chien and S.-W. Huang analyze structural context of catalytic residues based on theoretical and experimental structure flexibility. The results have shown that catalytic residues possess distinct structural features and contexts, and their neighboring residues within specific range are usually structurally rigid than those of noncatalytic residues.

In "*Simpute: an efficient solution for dense genotypic data*," Y.-J. Lin et al. compare the imputation performance among six various bioinformatics tools with data generated by randomly masking the genotype data from the International HapMap Phase III Project. They also propose a novel algorithm, Simpute, which is suitable and efficient for regular screening of the large-scale SNP genotyping in general.

In "*In silico prediction and in vitro characterization of multifunctional human RNase3*," P.-C. Lien et al. apply computational approaches for unique peptide extraction and perform *in vitro* activity assays for discovering important peptides in human ribonuclease 3 (hRNase3), HBP*rnase3*. They also identify multiple biological features of this unique peptide in glycan binding, cellular binding, and lipid binding, which are also characteristic features of hRNase 3. Their results demonstrate molecular evolution of sequence, structure, and function in human ribonuclease A (hRNaseA) superfamily members.

In "*Using nanoinformatics methods for automatically identifying relevant nanotoxicology entities from the literature*," M. García-Remesal et al. present a nanoinfomatics approach based on NER techniques for automatically identifying relevant nanotoxicology entities in scientific papers. The proof of concept can be expanded to stimulate further developments that could assist researchers in managing data, information, and knowledge at nanolevel and accelerating research in nanomedicine.

In "*On the difference in quality between current heuristic and optimal solutions to the protein structure alignment problem*," M. Arriagada and A. Poleksic utilize an approximation algorithm for protein structure matching to demonstrate that a deep search of the protein superposition space leads to increased alignment accuracy with respect to many well-established measures of alignment quality. The topic of protein structure alignment is still one of the most important problems in computational biology.

In "*Cancer vaccines: state of the art of the computational modeling approaches*," F. Pappalardo et al. introduce the new field of computational modeling of cancer vaccines which are a real application of the extensive knowledge of immunology to the field of oncology.

The papers in this special issue, representing a broad spectrum of computational approaches and areas of investigation, provide useful message of intelligent informatics for biomedical and biomedicine applications. This unique and informative collection of papers on bioinformatics highlights the direction of related studies. This special issue illustrates the importance that computational biology always plays the primary key step for biologists and medical doctors to be able to access large amounts of biological data.

Acknowledgments

Here, we want to thank the authors and reviewers for their scientific contribution and congratulate them for the high quality of their work.

Hao-Teng Chang
Raffaele A. Calogero
Sorin Draghici
Oliver Ray
Tun-Wen Pai

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 203681, 16 pages
http://dx.doi.org/10.1155/2013/203681

Methodology Report

Time Series Expression Analyses Using RNA-seq: A Statistical Approach

Sunghee Oh,[1] Seongho Song,[2] Gregory Grabowski,[1] Hongyu Zhao,[3] and James P. Noonan[3]

[1] *Department of Pediatrics, Children's Hospital Medical Center, Cincinnati, OH 45229-3039, USA*
[2] *Department of Mathematical Science, University of Cincinnati, OH 45221-0025, USA*
[3] *Department of Genetics, Yale School of Medicine, New Haven, CT 06520-8005, USA*

Correspondence should be addressed to Sunghee Oh; sunghee.oh@cchmc.org and James P. Noonan; james.noonan@yale.edu

Received 6 October 2012; Revised 10 January 2013; Accepted 15 January 2013

Academic Editor: Tun-Wen Pai

RNA-seq is becoming the *de facto* standard approach for transcriptome analysis with ever-reducing cost. It has considerable advantages over conventional technologies (microarrays) because it allows for direct identification and quantification of transcripts. Many time series RNA-seq datasets have been collected to study the dynamic regulations of transcripts. However, statistically rigorous and computationally efficient methods are needed to explore the time-dependent changes of gene expression in biological systems. These methods should explicitly account for the dependencies of expression patterns across time points. Here, we discuss several methods that can be applied to model timecourse RNA-seq data, including statistical evolutionary trajectory index (SETI), autoregressive time-lagged regression (AR(1)), and hidden Markov model (HMM) approaches. We use three real datasets and simulation studies to demonstrate the utility of these dynamic methods in temporal analysis.

1. Introduction

RNA-sequencing (RNA-seq) has fundamentally become the choice of studies of transcriptome [1–6]. From the conventional technologies in microarray and beginning of digital sequencing SAGE [7], a significant hurdle in the analysis of the transcriptome arises from insufficient samples, specifically, in identification of the temporal patterns of gene expression measured at a series of discrete time points. Several data-mining techniques and statistical methodologies have proven to be useful to search temporal gene expression patterns in microarrays [3, 8–34]. Some people have already started to adapt the way we applied in microarrays for RNA-seq data. The main drawback, however, is the loss of discreteness property of read count on transcriptional level, albeit there are no additional advantages in analytical aspects on counts. Given experimental design with sufficient replicate, time points, and sequencing depth [4, 5, 34], attempts to RNA-seq specific methodologies to preserve the elegant count property in time course will contribute to development and application in this area ahead. The last four years witnessed the astonishing publication of statistical methodology studies to identify differential expression between two or amongst multiple groups. Nonetheless most analysis tools remain tied to a static model approach without respect to time, albeit the incisive ultrahigh-throughput sequencing data now provides time series gene expression profile. As the first step towards understanding temporal dynamics in RNA-seq data, temporal analyses often rely on the simple pairwise comparisons [35–47] to infer differentially expressed genes/isoforms at a specific time point versus a reference time point. Differential expression results are then combined to characterize the dynamics over time. Commonly used microarray data analysis methods, such as limma [40], log linear models [39], and ANOVA [26], after variance-stabilizing transformation have also been used for temporal data analysis in RNA-seq as another alternative. However, the very few replications for such data limit the power of these methods. Statistical inference from such high dimensional data structure with the large number of variables and very few observations has presented substantial challenge. More importantly, the pairwise approaches fail to account for the strong temporal dependencies; indeed, higher correlation between neighboring time points is clearly revealed in published gene expression profiles [48] and our

real data applications (see Figure 2). Therefore, these pairwise approaches are suboptimal without explicitly modeling the expression dynamics over time nor can the time points that contribute most to time evolution trajectory pattern of a gene's expression be identified. More descriptive methods, such as clustering methods, have also been applied to identify coexpressed gene sets using RNA-seq data [49–51]. Such unsupervised clustering methods implicitly assume that data collected at different time points are independent, ignoring the sequential structure in time series data. It is apparent that potentially useful information on gene regulation and dynamics may be lost with these suboptimal methods, and there is a need to develop statistical methods that can appropriately model and analyze RNA-seq data. We discuss several methods that explicitly model the time-dependent nature of the time series data in this paper. We describe the identification of temporal differential expression (TDE) analyses as well as the ranking of genes to show temporal trajectories with statistical significance. We also discuss the application of time-lagged autoregressive AR(1) models to identify TDE genes as well as hidden Markov models (HMM) to classify different expression patterns by posterior probabilities of latent states. These methods can be applied to study complex factorial designs that interrogate multiple biological conditions simultaneously where multiple time points are studied under two or more biological conditions. Multivariate approaches are presented to identify temporal patterns in coexpressed gene groups and quantify coupled relationship of two distinct trajectories. Here we report an in-depth analysis of temporal patterns based on nonparametric and Bayesian approaches that incorporate the context of inherent time dependence of gene expression *per se*. When these methods are applied for published real datasets, both static and dynamic methods performed well for most temporal genes; however, dynamic methods had particularly a slight edge at low and moderate expression levels. That may be particularly advantageous for years to come for application to data with relatively low signals such as depression and aging data, which on expression compared to tumors in disease tissues.

2. Statistical Methods

2.1. Time Series Data Structure. Suppose that a gene expression profile matrix contains $i = \{1, \ldots, G\}$ genes and $j = \{1, \ldots, m\}$, m different time stages. The ith gene expression profile vector, $Y_i = [y_i(t_1), \ldots, y_i(t_m)]^t$,corresponds to a sequential vector of time points and biological replicates within a time point, namely, where $y_i(T = t_j) = [y_{it_jL=1}, \ldots, y_{it_jL=l}]$ is a vector composed of intraexpression measurements by $L = \ell$ biological replicates at time point $T = t_j$. We consider a sequence of observations on gene expression profile dataset, made at m different time points; accordingly m dimensional gene expression vector of gene i with observed read counts over time is used hereafter. $y_{ij} = [y_{ij1}, y_{ij2}, \ldots, y_{ijc}]^\tau$ is $C = c$ dimensional gene expression vector of gene i, time point j. The expression profile is a factorial time course experiment and the vector y_{ij} represents the intraexpression profile of c biological condition within a time point. $y_{ijc} = [y_{ijc1}, y_{ijc2}, \ldots, y_{ijcl}]^\tau$ is an l dimensional gene expression vector of gene i, time point j, c biological condition, and l different biological individual replicates. If there are not any treated biological conditions, the gene expression time series is simplified in y_{ijl}.

2.2. Statistical Evolutionary Trajectory Index (SETI). Existing static methods for testing significance of TDE genes in time series RNA-seq data do not consider temporal stochastic ordering dependency property in time, which differs from a typical gene expression profile data, and all static methods assume samples that are distributed independently and are not related to each other instead. However, it is well known that the considerable genes in gene expression profiles related to many developmental biological processes or disease progression are temporally differentially expressed and current expression level is affected by previous one by inherent Markovian property in time series. In the settings of large numbers of variables and with few observation, distribution-free or Bayesian approaches by using useful prior information are more suitable in RNA-seq. To circumvent the limitations and cope with a variety of particular patterns in time course, we present a statistical framework that enables more precise temporal expression profiling by incorporating autocorrelation measurement to determine relationship between consecutive expression profiles. Residuals in one period (or time point) are correlated with those in previous periods (or time point) and ranking individual SETI based on nonparametric regression fit as a gene-by-gene approach. As above, the gene expression level y_{ij} at $I = i$th gene, $J = j$th time, $C = \{1, \ldots, C\}$ biological condition, and $L = \{1, \ldots, L\}$ replicates is fitted by smooth spline regression. The autocorrelations of the residuals are computed by the sliding of all possible cases over the original time series, which are referred to as a trajectory index for given gene. The unbiased estimate of the autocorrelation for each gene is

$$\hat{\mathrm{ACR}}_{\mathrm{res},i}(k) = \frac{1}{(m-k)\,\sigma^2} \sum_{j=1}^{m-k} \left[Y_{ij} - \hat{y}_i\right] \left[Y_{ij+k} - \hat{y}_i\right] \quad (1)$$

for any positive integer $k < G$. $\{y_1, y_2, \ldots, y_G\}$ is a vector to be contained by G-length observations of expression measurement. P values for assessing statistical significance are calculated using a permutation test ($N = 10,000$), assuming the absence of temporal differential expression. The confidence interval and trimmed mean of trajectory index are derived by bootstrapping analysis ($B = 100$). The method is based on computing autocorrelations, that is, cross-correlation of gene expression profile across time points to represent temporal pattern. It is applied in a variety of different types of RNA-seq time series data including factorial time course experiments.

2.3. Autoregressive Time-Lagged AR(1) Model. We propose to use an autoregressive time-lagged AR(1) model for the identification of temporal and differential gene expression. Hay and Pettitt [54] demonstrated first-order time lag for an application to the control of an infectious disease with

count data over time in which the time series observations are examined to identify significant associations with explanatory variables and counts, the incidence of an infectious disease ESBL-producing Klebsiella pneumoniae in an Australian hospital, and the explanatory variable is the number of grams of antibiotic third-generation cephalosporins used over that time period. In order to essentially propose a universal dynamic method with AR(1) model in RNA-seq, we consider models to allow flexibility without covariates in lieu of taking their initial approaches. The details of our AR(1) model for read count gene expression profile over time as a gene-by-gene TDE identification are discussed in the following with mathematical notations. Bayesian framework is defined by $(y_{ij} \mid \mu_{ij}, i = 1, \ldots, n$ and $j = 1, \ldots, m)$ to be independently distributed as Poisson model. We employ their model for RNA-seq read count expression data.

2.3.1. Poisson Model in AR(1). From the time series data structure (Section 2.1), we have m time points, c biological conditions, and l replications. Both maize and zebrafish data with single measurements within a time point are applied in this method.

Consider

$$\begin{aligned} y_{ij} &\sim \text{POI}\left(\mu_{ij}\right), \quad \text{where } i = 1, \ldots, n, \ j = 1, \ldots, m, \\ \log\left(\mu_{ij}\right) &= w_{ij} + \beta_i, \\ w_{i1} &= \frac{u_{i1}}{\sqrt{1 - \varphi_i^2}}, \\ w_{ij} &= \varphi_i w_{ij-1} + u_{ij}, \quad j > 2. \end{aligned} \tag{2}$$

And equivalently,

$$\begin{aligned} \log\left(\mu_{ij}\right) &= w_{ij} + \beta_i, \\ w_{i1} &\sim \text{Normal}\left(0, \frac{\sigma^2}{\left(1 - \varphi_i^2\right)}\right), \\ w_{ij} \mid w_{i,1\ldots,j-1} &\sim \text{Normal}\left(\varphi_{i1} w_{ij-1}, \sigma^2\right), \quad j > 2. \end{aligned} \tag{3}$$

To identify altered gene expression across time series, for each gene, the AR(1) model is applied and inference of β is obtained from noninformative priors and time series random effects for sequential expression profile are assumed. This autoregressive model was originally carried out for longitudinal large-scale historical repeated-measurement data. In our study, using the modified assumptions, RNA-seq time series with short time period (4~8 time points) and single observations as gene-by-gene approach are applied to compare the performance of AR(1) model to static methods in identification of differential expression. The posterior probabilities of parameters in the model are estimated through MCMC simulations with $N = 6{,}000$ iteration and 1,000 burn-in. We provide detailed notations and equations for three dynamic approaches in Supplementary data available online at http://dx.doi.org/10.1155/2013/203681. In the results, we are most interested in autocoefficients to represent time series sequential random effects in the model and we implemented a classification between TDE (temporally differential expression over time) and EE (equally expression over time) set of genes. Similar to statistical differential expression testing for each gene in a classical approach, our implementation of testing in AR(1) model is given by a Bayesian interval estimate, 95% credible interval:

$$H_0 : \text{if } \varphi_i = 0, \text{EE}, \qquad H_1 : \text{otherwise, TDE}, \tag{4}$$

where we consider that gene i is temporally differentially expressed (TDE) if the 95% credible interval of φ_i does not include 0; otherwise it is considered to be equally expressed (EE). Also we obtain the tail probability of $(\varphi_i \mid y)$ of gene i, that is, $p(\varphi_i > 0 \mid y)$ or $p(\varphi_i < 0 \mid y)$ for $i = 1, \ldots, n$ using MCMC. It indicates the significance of differential expression for each gene.

2.3.2. Negative Binomial in AR(1). A more compelling methodological goal is to infer temporal dynamics when we have replicates within a time point and it is straightforward to establish a negative binomial model with AR(1):

$$\begin{aligned} y_{ij} &\sim \text{NBC}\left(k, \mu_{ij}\right), \quad \text{where } i = 1, \ldots, n, \\ & j = 1, \ldots, m, \ \log\left(\mu_{ij}\right) = \omega_{ij} + \beta_j. \end{aligned} \tag{5}$$

Other parts of the model remain identical as in (3). Here, $y \sim \text{NBC}(k, \mu)$ means that y has its probability function as follows:

$$\begin{aligned} p\left(y; k, \mu\right) &= \frac{\Gamma\left(y + k\right)}{\Gamma\left(k\right)\Gamma\left(y + 1\right)}\left(\frac{k}{\mu + k}\right)^k \\ &\quad \times \left(1 - \frac{k}{\mu + k}\right)^y, \quad y = 0, 1, 2, \ldots. \end{aligned} \tag{6}$$

This negative binomial distribution has its mean $E(y) = \mu$ and its variance $\mu + \mu^2/k$. The parameter k^{-1} is called the dispersion parameter.

2.4. Hidden Markov Model (HMM). We consider a Bayesian HMM to analyze factorial time course RNA-seq data. Our model follows the seminal work of Yuan and Kendziorski [48] that characterizes all possible temporally differential expression patterns in time series microarray data with two or more biological conditions. Although this early study was encouraging, the HMM was restricted to represent timing differences between biological conditions with binary EE/DE or multiple cases of latent hidden states depending on the number of given conditions at each time point. The extent of temporal changes was not obvious in significantly differentiating between one time point and the next. Taking a HMM approach, we seek SETI and multivariate coupled relationships among distinct trajectories into HMMs in each condition to investigate biological evolutionary trajectory that can be applied to a comprehensive set of RNA-seq time series data to make probabilistic predictions of temporal patterns for how differential expression will occur under different biological conditions. Also, count specific underlying

distributions for RNA-seq time series data are used. First, we introduce a mechanism to use the inference of temporally differentially expressed genes in time series RNA-seq gene expression profiles with multiple biological conditions at a given time point. This was achieved by incorporating GP and NBD with corresponding prior information into the HMM for each gene, allowing samples having either multiple replicates or single observations. We investigate properties of the HMM technique such as how it benefits by incorporating hidden variables when making the predictions of temporal patterns of differential expression for given different biological conditions and how the number of chosen latent variables varies with conditions within a stage over a time period. As per Section 2.1, we present how to express hidden states in the given models with subindices composed of T time points, C different biological conditions (e.g., drug treatments or tissues), and L replicates. As RNA-seq experiments generally have small sample sizes, the identification of statistically significant temporally differentially expressed (DE) genes may have limited power. Also, some studies stress the importance of replication in microarray studies, which have inherent variability [4, 5, 33, 34, 55] regardless of how well constructed DE methods are applied. Thus, without replicates, no statistical significance tests are reliable and powerful on detection of TDE. With the reduction in sequencing costs, well-designed balanced RNA-seq experiments with proper sample sizes and time points will facilitate the use of temporal dynamic methods, including AR(1) model. Here HMM is used with samples and 4 biological conditions (different tissues). Consider that the gene expression dataset (y_{ijcl}) has I = genes, J = time points, C = conditions, and L = replicates. This algorithm has the Markovian assumption that the expression level at the current time only depends on that at the most recent time. We use hidden states to represent a change in expression levels between different biological conditions. Thus, this framework allows us to detect TDE genes and to facilitate the calculation of the posterior probabilities of all possible TDE patterns. For instance, with three time points, this method can estimate the posterior probability of pattern EE-DE-EE, where EE stands for equally expressed and DE for differentially expressed, respectively. Namely, the main interest is to identify the relationship among the C class latent mean values of expression level for each gene g at each time point $T = t$ denoted by $\mu_{gt1}, \mu_{gt2}, \ldots, \mu_{gtC}$. Hereby, the primary goal of HHM in time course experiment with multiple different conditions is to infer all potential relationships from different conditions; for simplest case with two biological conditions, it is binary outcome with EE/DE, and for complicated experimental design with more than two biological conditions, suppose that biological conditions correspond to different tissues, hereafter tissues A, B, C, and D. Correspondingly, there are 4 expression profiles $\mu_{gt\mathrm{A}}, \mu_{gt\mathrm{B}}, \mu_{gt\mathrm{C}}$, and $\mu_{gt\mathrm{D}}$, and 15 possible expression pattern states include the following:

$$\begin{aligned}
&\text{State 1 [1111]} : \mu_1 = \mu_2 = \mu_3 = \mu_4\\
&\text{State 2 [1221]} : \mu_1 = \mu_4 \neq \mu_2 = \mu_3\\
&\text{State 3 [1222]} : \mu_1 \neq \mu_2 = \mu_3 = \mu_4\\
&\text{State 4 [1121]} : \mu_1 = \mu_2 = \mu_4 \neq \mu_3\\
&\text{State 5 [1212]} : \mu_1 = \mu_3 \neq \mu_2 = \mu_4\\
&\cdots\\
&\text{State 14 [1233]} : \mu_1 \neq \mu_2 \neq \mu_3 = \mu_4\\
&\text{State 15 [1234]} : \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4
\end{aligned} \tag{7}$$

More generally, the number of all potential patterns as a function of the number of tissues is equal to the Bell exponential number of possible set partitions. Here each state is not observed and needs to be estimated from the data. Therefore, we refer to such states as hidden. For each gene g at each time point $T = t$, we want to estimate the probability of each hidden state $p(\vec{g}_{gt} = k)$ and then we associate an observation model with each state and eventually also compute the most likely sequential states over time to derive timing differences for a given gene g. Fitting a hidden Markov model involves estimating the transition probability matrix A, initial probability distribution π_0, and unobserved hidden state at time $T = t$, and estimations are done by EM algorithm as described and implemented in the original paper of HMM. The parametric empirical models (PEM) of GP and NBD sample $y = (y_1, y_2, \ldots, y_N)$ are considered here.

In the GP model, for two biological conditions at each time point and two marginal distributions of hidden states are given the following equations, as shown in Yuan et al., for microarray application. The underlying distributions and joint predictive density (JPD) for discrete count data are incorporated to infer posterior probability distributions:

$$f_{1t}\left(x_{gt}\right) = f_{0t}\left(x_{gt} \mid \lambda_{gt}\right) dG_t\left(\lambda_{gt}\right) \tag{8}$$

under EE state 1 and

$$\begin{aligned}
f_{2t}\left(x_{gt}\right) = &\int f_{0t}\left(x_{gt1,\ldots,gtn_1} \mid \lambda_{gt}\right) dG_t\left(\lambda_{gt}\right)\\
&+ \int f_{0t}\left(x_{gt(n_1+1),\ldots,gt(n_1+n_2)} \mid \lambda_{gt2}\right) dG_t\left(\lambda_{gt2}\right)
\end{aligned} \tag{9}$$

under TDE state 2.

If π_i represents the proportion of TDE genes at time t, then the mixture type of marginal distribution of the data is given by

$$\left(1 - \pi_1\right) f_{1t}\left(y_{gt}\right) + \pi_2 f_{2t}\left(y_{gt}\right), \quad \text{where } i = 1, \ldots, d. \tag{10}$$

And $f_{ot}(y \mid \mu_{gt}) = \lambda \exp(-\lambda y)/y!$, $x > 0$. λ_t follows a conjugate prior with gamma distribution parameters, shape parameter α_t, and rate parameter β_t. Thus, three parameters $\theta_t = (\lambda_t, \alpha_t, \beta_t)$ need to be estimated for a given gene. For the GP model, the Markov chain is assumed to be homogeneous and the marginal distribution of x_{gt} is the finite mixture $\sum_{i=1}^{d} \pi_i f_{it}$. We assume one-step first-order correlation time series structure so that HMM contains with

Poisson distributed state-dependent distribution. The goal of this algorithm is to identify a certain set of genes that are TDE in a combination of time series and four different biological conditions, for example, distinct tissue types. To address the utility of HMMs proposed in time course RNA-seq experiments with multiple different tissues, we exploit a parametric hierarchical empirical Bayes model with GP (data w/o replication) and NBD (data w/replications) with beta-prior as a well-modified Bayesian approach [42, 56, 57]. The Newton et al. [57] approach identifies differentially expressed genes for microarray experiment framework in multiple biological conditions at a static time point and similarly Hardcastle and Kelly [42] identify differentially expressed genes either for pairwise comparisons or for multiple group comparisons in an RNA-seq experiment framework at a static time point. For microarray data, Yuan and Kendziorski [48] proposed a HMM for a dynamic time course experiment with multiple conditions Gamma-Gamma (GG) and Log Normal Normal (LNN) to identify genes of interest whose temporal profiles are different across two or more biological conditions. Here, we adapted and extended that approach to a general RNA-seq framework with GP and NBD models as more flexible models. The earlier studies are limited to detect temporal patterns other than ranking/ordering temporal dynamic specific genes during developmental stages, which biologists are more interested in examining. We assume two common underlying distributions for RNA-seq read count. In reality, violation of GP assumptions is very common and in order to account for overdispersion. Alternatively, NBD is applied with a beta-prior. The above inference method provides for continuous trajectory regression involved with timing evolution features to rank temporal genes statistically for a given pattern, as well as such genes' temporal differential expression patterns among conditions. In addition, we examined multivariate identification of temporal expression using the following several metrics.

2.5. Coupled Multivariate Identification of SETIs

2.5.1. Granger Causality. The concept of Granger causality between two distinct SETIs assumes that the data at the current time point affect the data at the succeeding time point [58]. To determine Granger causality for each pair of trajectories, we employ standard F-statistics to test if the residual values derived from the fitting smoother for gene A are incorporated into the equation for another gene B. If all the coefficients for the measurements of gene B are zero under the null hypothesis, then there is no statistically significant Granger causality between the trajectories for genes A and B.

2.5.2. Cotrajectory with Glass-d-Score. Similarly, each pair of two trajectories, which correspond to two gene expression levels, is explored by another dependency metric score and detailed notations are described in the following, when there is a given pair of two gene expression profiles:

$$\left(g_i, g_j\right) d_k^{ij} = \frac{r_k^{ij} - \overline{r}_k^i}{\sigma_{r_k^i}}, \tag{11}$$

where r_k^{ij} is the correlation coefficient between the expression profiles of (i, j) among all possible pairs. The null distribution was assumed to have $\overline{r}_k^i(\sigma_{r_k^i})$ the mean and standard deviation of correlation coefficient between gene i and all other genes, respectively.

2.5.3. Correlation Approach. As proposed in Ma et al. and Barker et al. we propose a biologically motivated approach to measure the relationship between two different genes based on their temporal expression profiles in RNA-seq. Ma et al. proposed to consider lagged coexpression analysis to capture the scenario that there is a delayed response of gene B to gene A so that the profile of gene B is correlated with the time delayed profile of gene A.

2.6. Pairwise Methods.

In this section, we describe the pairwise methods that we consider in our comparisons with the methods discussed above that can explicitly model the time dependencies nature in the data. For comparisons with our dynamic methods, we examined several popular static methods, including Fisher's exact test for simple two sample comparisons and log linear model for multigroup comparison, which can also be applied for RNA-seq time series data in temporal analysis as intuitive but limited.

DE analyses: we first employed pairwise condition comparison methods in digital measures at a given static status without respect to time. It is no surprise to take a union set of all possible pairwise comparisons using these static techniques to identify temporal dynamics in relatively small experiments, where single sample for each time point and very few number of time points are contained in experimental design.

(i) Fisher's exact test: from Table 1, the 2-sided P value for TDE of each gene is computed with (12) [39]:

$$\Pr\left(g_{+1,g} = g \mid g_{+1.}, g_{+2.}, \ldots, g_{.g}\right) = \frac{\binom{g_{+1.}}{g}\binom{g_{+2.}}{g_{.g}-g}}{\binom{g_{..}}{g_{.g}}}. \tag{12}$$

(ii) Audic-Claverie statistics.

The Audic-Claverie statistics [59] are based on a distribution $p(y \mid x)$ over read counts y in one sample in one given group informed by the read counts x under the null hypothesis that the read counts are generated identically and independently from an unknown Poisson distribution. $p(y \mid x)$ is computed by infinite mixture of all possible Poisson distributions with mixing proportions equal to the posteriors under the flat prior over λ. When the two libraries in a given Solexa/Illumina RNA-seq experiment are of the same size,

$$p\left(y \mid x\right) = \frac{1}{2^{x+y+1}} \frac{(x+y)!}{x!\,y!} = \frac{1}{2^{x+y+1}} \binom{x+y}{x}. \tag{13}$$

These are Audic-Claverie statistics [59] for given read counts x and y.

Pooling methods: as with ANOVA in microarray, log linear model and linear models for microarray data (LIMMA),

Table 1: 2 × 2 contingency table.

	Reads from sample of type		
Tags in gene	Group A	Group B	
Gene g	$g_{+1,A}$	$g_{+2,A}$	$g_{\cdot,A}$
Not gene g	$g_{+1} - g_{+1,A}$	$g_{+2} - g_{+2,A}$	$g_{\cdot,B}$
Total	g_{+1}	g_{+2}	$g_{\cdot\cdot}$

after variance-stabilizing transformation to allow multigroup and multifactor comparisons, can be applied by including a time variable as the main factor in the model [40].

(i) Log linear model with the Poisson link function (or negative binomial when replicates are available) and likelihood ratio test model. In the model, the time factor, biological condition factors, and their interaction terms are included.

(ii) LIMMA (linear model for microarray) with F-statistics under the linear model setting implemented in R package is also applied for time series RNA-seq read count data after variance stabilizing transformation.

Although such static algorithms have demonstrated a successful identification of temporally expressed genes in some degree in the past four years and our study, any temporal dynamic analysis false discovery results in static methods can be introduced due to violation of Markovian assumptions frequently revealed in time series expression profile. As the cost to sequencing continues to decline, there is urgent need for more sophisticated statistical methodologies of power in the identification of temporal expression or for use of characterization of temporal dynamics to assess isoform diversity within a gene level in a future investigation of time series RNA-seq. Ideally, it is very critical to appropriately have a good model to represent observed data since interpretation of a model that does not contain valuable information is useless. For this important purpose, our dynamic methods are compared to these static methods by evaluating the overlap in the number of differentially expressed genes in real data sets.

3. RNA-seq Time Series Data

3.1. Three Different Types of Time Series. There are mainly two types of time series in RNA-seq. The first is factorial time series data that include at least two biological conditions to be compared in a given time point and have multiple developmental patterns over time as the number of conditions. The second type of time series has a single condition and corresponding developmental stage. In the third type of time series, there are subsequently two additional subtypes, circadian rhythmic data and cell cycle data. In this study, we formulate the statistical framework of identification of temporal changes in RNA-seq time series for first two types of data and the periodic data-sets are reviewed in "another review manuscript" with discrete Fourier transformation and other methods in a separation in depth.

3.2. RNA-seq Real Time Series Datasets

3.2.1. Factorial Time Course Experiment: A Sheep Model for Delayed Bone Healing. We consider this published RNA-seq time series data from a sheep model for delayed bone healing. In Jager et al., surgery was conducted as described in [52, 53] and the newly generated tissues were harvested at different days, 7, 11, 14, and 21 after surgery. For each time point, there are 6 biological replications for both groups except one time point, for day 21 (group I, $n = 5$, group II, $n = 6$), where two groups are defined by standard healing system and delayed healing system. Thus, the authors considered two treatments: standard healing system and treatment with unstable external fixator leading to delayed bone healing. While the standard bone healing system was investigated in a 3 mm tibial osteotomy model stabilized with a medially mounted rigid external fixator, delayed healing was investigated in a 3 mm tibial osteotomy model stabilized with a medially mounted rotationally unstable external fixator. For each treatment, RNA-seq data were collected at 4 time points: 7, 11, 14, and 21 days, with 5-6 individuals' DNA samples pooled together at each time point. In their differential expression, they used the pooled samples from 5~6 lanes of animal samples at one time point and Audic-Claverie statistics were performed using 4 samples over 4 time points by taking a union set of all possible pairwise comparisons using static methods. We reanalyzed their sheep animal time series data using three dynamic methods to identify TDE genes.

3.2.2. Single Transient Time Course Experiment-I. We applied two single biological condition time series data which are interested in exploring developmental transient patterns during a time period rather than timing difference patterns incorporated with multiple conditions at a time as Section 3.2.1 example. Maize leaf transcriptome with four different developmental zones containing two replicates in each time point [50] was employed. This is one representative for time course experiment with single transient expression profile. Tissues were collected from leaf 3 at 9 days after planting 3 hours into the L period from four segments: (1) basal (1 cm above the leaf three ligule), (2) transitional (1 cm below the leaf two ligule), (3) maturing (4 cm above the leaf two ligule), and (4) mature (1 cm below the leaf three tip). Thus, maize leaf data with different developmental stages are generated from mRNA isolated from four developmental zones: basal zone, transitional zone, maturing zone, and mature zone. In the differential expression analysis, they simply applied chi-squared static method and K-means clustering method that both do not take into account time dependency, but all samples are assumed to be independent. This maize leaf time series data are reanalyzed with proposed methods in this study.

3.2.3. Single Transient Time Course Experiment-II. This is a time series experimental design to be composed of eight stages during early zebrafish development, embryogenesis [51]. In their study, wild-type zebrafish embryos (TLAB) were

Table 2: Statistical evolutionary trajectory index (SETI) of the top candidate genes where FDR is controlled at less than 0.05 in a sheep model data. The gene expression level is fitted on smooth spline function, autocorrelation of residuals is measured, and corresponding statistical significance is tested. In addition, trimmed mean of bootstrap and 95 percent of confidence interval are also provided in the table.

Top candidate genes	SETI (trimmed mean of bootstrap)	Bias of bootstrap	95% CI of SETI	*P*	FDR
A5D9H5	1.23 (1.23)	0.12	[1.02, 1.43]	0	0
A6QQB6	1.23 (1.27)	0.09	[1.00, 1.46]	0	0
A0JN96	1.23 (1.23)	0.09	[1.03, 1.43]	0	0
DUFFY	1.23 (1.23)	0.13	[1.02, 1.43]	0	0
A6QP68	1.23 (1.23)	0.10	[1.03, 1.42]	0	0
gi|11992112	1.23 (1.23)	0.10	[1.05, 1.40]	0	0

staged according to standard procedures and about 1,000 embryos were collected per stage (two to four cells, 1,000 cells, dome, shield, bud, 28 hpf, 48 hpf, and 120 hpf) within a tight time window of ~10 min. Their collection of embryos was ensured that all embryos were at the same developmental stage. The identification of long noncoding RNAs (lncRNAs) expressed during zebrafish embryogenesis was explored to assess a diversity of transcripts that are structurally similar to, but noncoding, mRNAs. The analyses of RNA-seq time series expression profiles focused on the identification of temporal dynamics of lncRNAs using the Cuffdiff method in its time series mode with upper quantile normalization, which is also limited to pairwise comparison from previous time point to right next time point. Here, the data reanalyzed the transcriptomic gene expression profile data with 28,520 annotated protein coding genes. To consider the possibility of similarities and differences in comparisons between static and dynamic methods for time series RNA-seq data, we systematically compared both methods with these data.

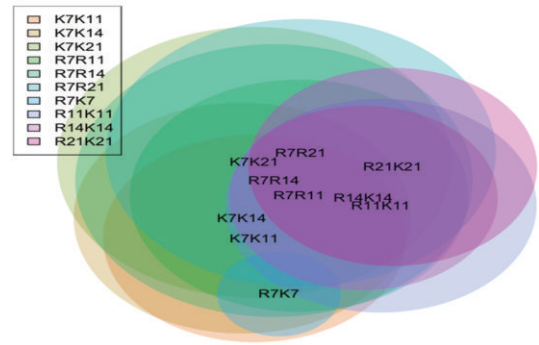


Figure 1: Venn diagram for differentially expressed gene sets detected by Fisher's exact test where the Benjamini-Hochberg FDR is controlled at <0.05. In this figure, as the labels authors used in [52, 53], K represents standard and R represents delayed healing system in a sheep model for two different bone healing systems.

3.3. Results in Differential Expression Analysis on Static and Dynamic Methods. For the sheep data, the authors applied the Audic-Claverie method to the normalized expression values, RPKM, to compare later time points to the reference time point (7 day) in both groups. After all pairwise comparisons, they combined the sets of differentially expressed gene sets with 884 genes detected in total from 24,325 mappable genes. Based on these 884 genes, they performed hierarchical clustering to identify gene clusters. Each cluster was then subject to gene ontology analysis to find significant biological functions. The differential analysis performed in original paper is based on static differential analysis method. We reanalyzed their sheep factorial time course experiment data to identify TDE genes over time through dynamic methods, HMM, SETI, and AR(1) model to account for correlated time-dependency structure. HMM identifies temporal patterns with classification of DE/EE at each time point by posterior probabilities, whereas SETI with statistical significance from permutation resampling procedures and AR(1) model with gamma Poisson Bayesian assumption on count data are applied within single biological condition, separately. Results obtained by these dynamic methods compared those of static methods, simple pairwise methods, Audic-Claverie statistics and Fisher's exact test, and pooling static methods, glmFit in edgeR, LIMMA, and log linear model as shown Figure 3. To identify temporal dynamics by assuming correlated data structure, we performed HMM modeling with Poisson-gamma since there were no replicates. AR(1) model and SETI significance tests were also done within each biological condition. Temporally differential expression gene sets detected by these dynamic methods were compared with the results of simple pairwise tests and pooling methods. From the HMM, 646 temporal dynamics of DE calls are identified to represent DE in at least one time point. The HMM model only explores different temporal patterns of DE/EE states and does not rank the genes by statistical significance, but is classifying gene expression profile into a number of temporal patterns by posterior distribution of latent states. Because of this limitation, we

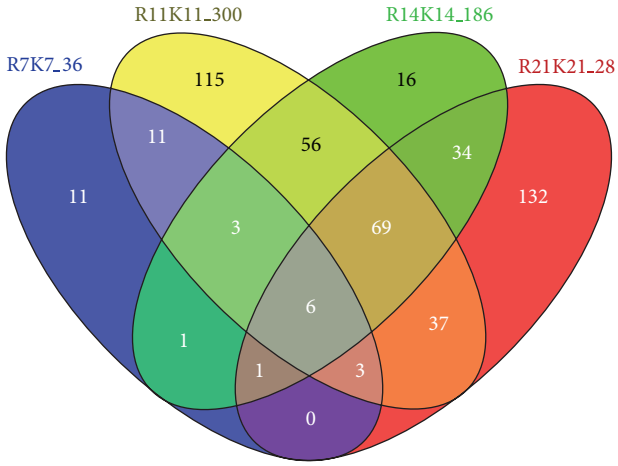


Figure 2: Venn diagram of four DE sets having the number of DE genes between two different healing systems detected by Fisher's exact test with FDR 0.05 at each time point ($t = 7, 11, 14$, and 21 days). Four time points were compared in simple pairwise comparison between two biological conditions, R (delayed healing) versus K (standard healing system). The label of each set depicts the number of DE genes in the specific comparison. The majority of interaction sets of DE genes between two successive time points implies that high proportion in detected differentially expressed genes at current stage tends to be redetected at next stage revealed by inherent time-dependent structure in time series gene expression profile.

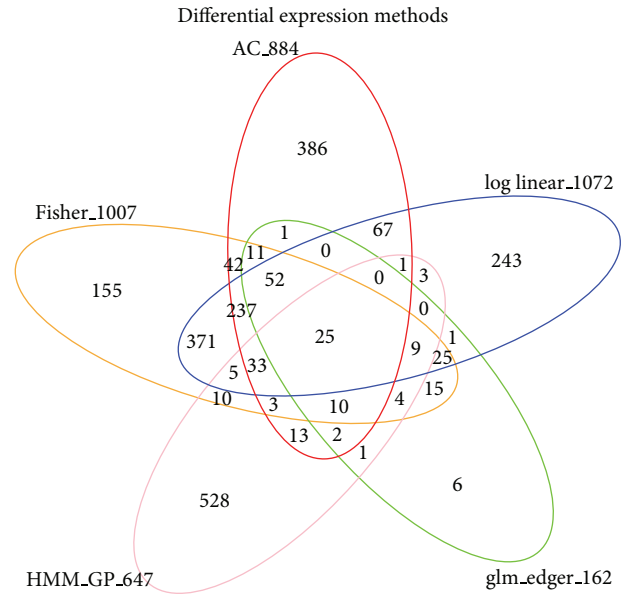


Figure 3: Venn diagram of five DE sets for static and dynamic methods in a sheep model data. In static methods, for two simple pairwise methods, Audic-Claverie statistics and Fisher's exact test were performed and both methods take a union set of all possible pairwise comparisons to identify temporally differentially expressed (TDE) genes across time points and two healing systems. As another static approach, pooling methods of samples, log linear model in [39], and generalized linear model fit in edge R in [36] were performed and detected TDE genes by FDR 0.05. In dynamic HMM method, we identify top candidate TDE genes defined at least showing DE pattern from one time point based on posterior probabilities for latent variables (DE/EE between given biological conditions). On the basis of comparison of the number of DE genes identified by each method, patterns of identification of TDE genes are method specific suggesting validation procedures of methods in biological aspects.

employed the SETI and AR(1) models to discover developmental transient patterns in each condition. The trimmed mean time evolution trajectory index is presented for the top three candidate temporal genes in each bone healing system. The 95% confidence interval of bootstrapping and FDR of permutation re-sampling are shown in Figure 4 and Table 2. To determine temporal dynamics and meaningful biological functions, only HMM-specific TDE genes which are not contained in static methods are further explored in gene clustering and biological functional network analysis as shown in Figures 5(a) and 5(b), respectively. In the results, they obviously showed temporally differential expression implying that loss of information to assumption of stochastic time-dependent structure might lead to false discoveries and less power of detection. To discover temporal transient patterns of differential gene expression within each biological condition; healing system, we performed SETI and AR(1) model approaches for each condition, SETI results are given in Figure 4 showing top candidate TDE genes, of which some genes such as gi|119921123 and B6DXC7 are of low expression levels which we were not able to detect in static methods. In the second data for our study, we have reanalyzed maize leaf transcriptome data to identify TDE genes with static and dynamic methods and compare between two. In their paper, they investigated leaf development gradient in time series gene expression data at successive stages (4 time points: base, tip: basal, transitional, and maturing) and identified a gradient of gene expression from base to tip: basal (23,354) > transitional (22,663) > maturing (22,036) > mature (21,332) from a total of 25,800 annotated genes. In the differential analysis in times series RNA-seq data, they used the method proposed in Marioni et al. [40] for pairwise analysis. A total of 16,502 genes were found to be differentially expressed in at least one of the comparisons. They then performed K-means clustering and showed eighteen clusters along the four developmental zones (Base, −1 cm, 4 cm, Tip). To compare gene sets detected by our dynamic methods with their gene lists, dynamic methods, SETI, and AR(1) model are applied again in this study and all temporally differentially expressed genes are presented in Supplementary Tables 1 and 3, where filtered gene set to be tested in differential expression has 5273 and 12,322 temporal dynamic transcripts from 42399 transcripts through SETI and AR(1) model, respectively. On the basis of significant temporal expression, we compared dynamic methods to static methods, which were used in the original paper without accounting for correlated data structure type. As the third real data application, to identify temporal dynamics, we have reanalyzed the third data, zebrafish embryonic transcriptome, focusing specifically on

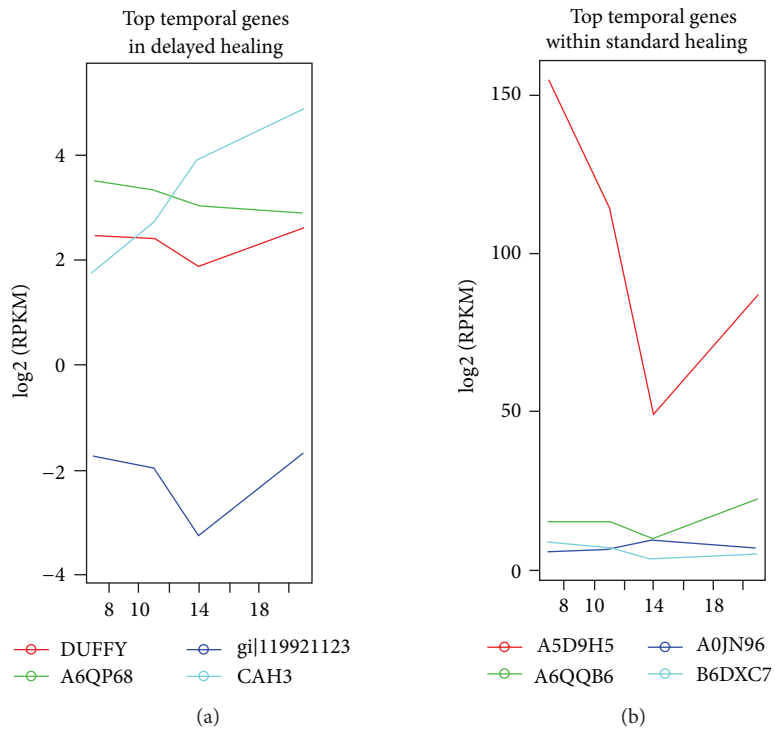
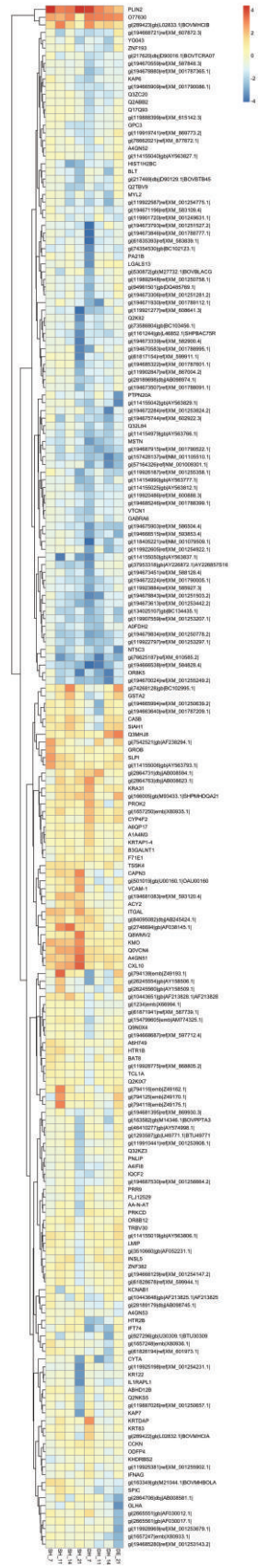


Figure 4: Top candidate temporally differentially expressed genes identified by statistical evolutionary trajectory index (SETI) within each healing system in a sheep model data. Each panel depicts temporal patterns between log2 (normalized expression levels, RPKM) and four different time points under their expression curves. The distinct colors represent significant individual genes ranked by SETI and FDR 0.05 by resampling procedures.

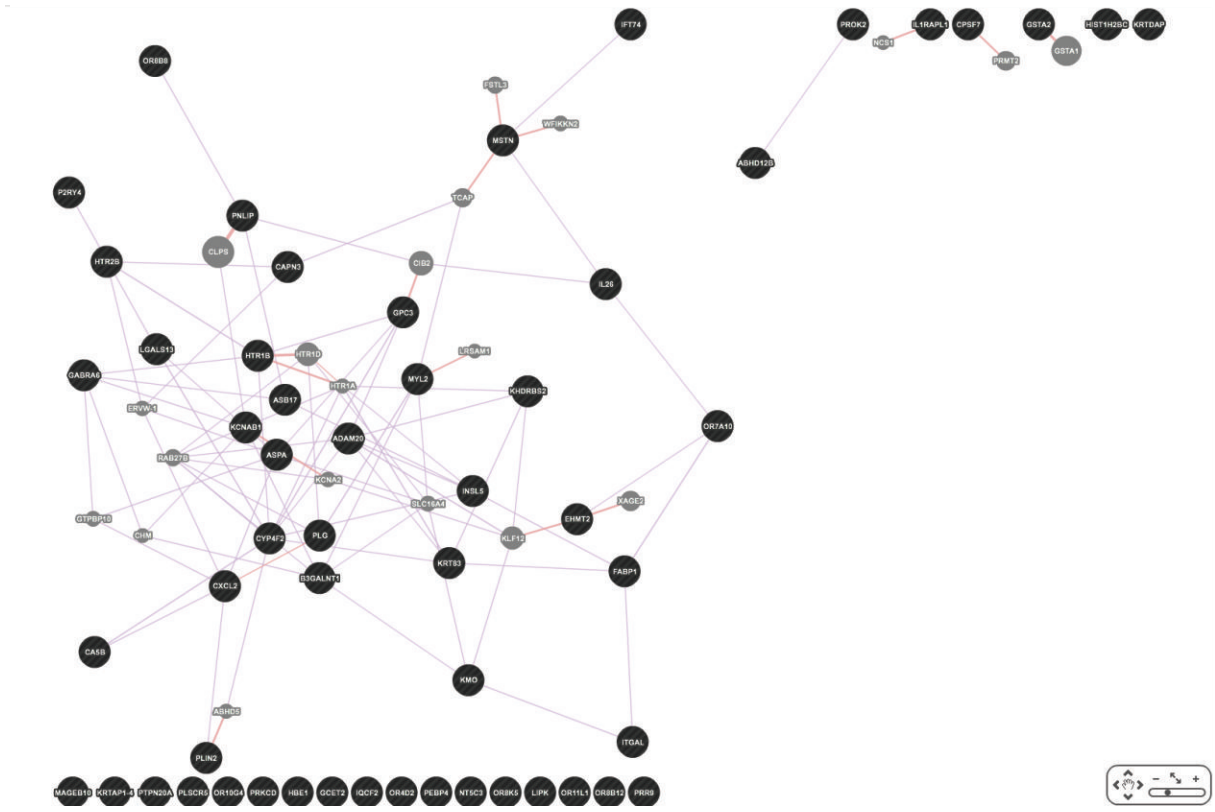
the identification and characterization of temporally differential expression using statistical evolutionary trajectory index and autoregressive time-lagged AR(1) model. We furthermore implemented both methods to rank temporal genes by statistical significance. As consequence of the resampling-based procedures and posterior probabilities of autocorrelation, it was possible for gene-by-gene approach to order temporal genes by two dynamic methods and identify genes associated with cotemporal dynamics. To investigate such paired temporal dynamics, we examined the relationships between genes using bivariate identification methods. Glass-s-d score is reported in Supplementary Table 5. Likewise, the statistical evolutionary trajectory index with statistical significance for zebrafish data is given in Supplementary Table 2, where we filtered out genes by coefficient of variation (CV) criteria remaining 12,034 genes. Overall, both methods show more robustness at low and moderate expression levels when compared to existing parametric static methods indicating that our methods achieve relative improvements in test of identification of temporal genes and AR(1) model shows more sensitive TDE calls than SETI resampling procedure in two real data applications. Here, we examined how different results are obtained by dynamic time series methods. For simple pairwise static methods, we employed Audic-Claverie statistics and Fisher's exact test as these two methods have been widely used in previous studies. They showed highly concordant results on other RNA-seq datasets compared to DEGseq, DESeq, edgeR, and baySeq (data not shown). In differential analysis with simple pairwise methods, we took a union set after all pairwise comparisons across a time period and amongst different biological conditions as these methods only consider two pairwise comparison testing and confirm the results to those of original papers. For pooling static methods, LIMMA, log linear model, and edgeR R package with glmFit are carried out to identify TDE genes. To compare with above static methods, we employed three dynamic methods described in the previous sections. The results are shown in Figures 1 and 3. Figure 2 shows how dependent structure is observed in patterns identified across time points, 36(23), 300(277), and 186(134), of the previous TDE gene set, genes in 64% ~92% percentage are differentially reidentified at the right next time point, implying that there is temporal dependent structure in sheep healing system RNA-seq time series data.

3.4. Bivariate Dynamic Analysis for RNA-seq Time Series. In systems developmental biology where characterization of complexity of various time course data likely leads to address inference of temporal dynamic patterns from transcriptome, we are not often really interested in exactly how only a single gene is temporally differentially expressed at a particular time point or period. This knowledge would neither answer an understanding of how biological networks in temporal dynamics of gene regulation work nor enable predicting any cooperative sets of genes to occur under biological



(a)

Figure 5: Continued.



(b)

Figure 5: (a) Coexpression patterns from gene clustering in a sheep model data. 200 HMM-specific TDE genes are represented in heatmap. Each row contains a vector of time series expression profile in log2 scale; consequently the visualization in heatmap is originally made up of major three groups, high, moderate, and low expression levels with genens that are not detected by static methods but detected by HMM, of which we selected the most statistically significant 200 genes to present this heatmap. Interestingly, some genes at low expression levels were obviously differentially expressed at log2-scaled FC ~4 up to 5 and even some genes that significantly show temporal patterns at high expression levels were also detected, yet those genes were not detected by existing static methods suggesting that HMM method reassuringly has higher sensitivity and robustness than other existing static methods in identification of differential expression regardless of expression levels. (b) Gene functional pathway and network analysis with 528 HMM specific TDE genes in a sheep model data. To explore biological functions in this gene set further, whether or not those are genuinely differential expression or random noise by chance in terms of biological insights, gene ontology (GO) and KEGG pathway analysis were performed to identify meaningful functionalities and some meaningful functions related to developmental process (intermediate mesoderm formation, regulation of cell growth involved in regulation of muscle adaptation, intermediate mesoderm formation, etc.) and gender specific terms (granulosa cell development and maternal placenta development) are detected as we anticipated to confirm the sensitivity of dynamic HMM method. The purple and pink legends represent coexpression and physical interactions across genes, respectively, and black nodes are query genes in networks.

conditions across time points. Thus, it is well known that genes work collaboratively together in a structured biological network; these biological phenomena underscore the importance taking into account the multivariate techniques when modeling temporal dynamic gene expression. Since it is not known beforehand which gene features are connected to each other, investigators sought to define informative relationships between individual gene patterns to identify many relevant classes of dynamic temporal gene expression patterns. We explored highly correlated relationships between temporal gene sets detected by bivariate dynamic methods. Pairs of trajectories were further investigated to explore the coupled coordinated relationships between different temporal patterns based on the three dependency metrics in Section 2.5. Significance levels of such relationships were estimated by bootstrapping resampling. The methodologies to test any coupled relationship to pairs of district gene expressions are based on (1) Granger causality, (2) correlation-basis approach, and (3) Glass-s-d score as defined in [60]. In order to efficiently identify copaired temporal dynamics, using zebrafish data, we first identified statistically significant TDE genes and ran Glass-d-score based on gene permutations. Figure 8 demonstrates coupled temporal dynamics with log-scaled expression level.

3.5. Gene Functional Pathway and Network Analysis. Once temporal dynamics in gene-by-gene test and in gene-to-gene interaction were determined, the resultant temporal gene expression sets detected by ranking individual

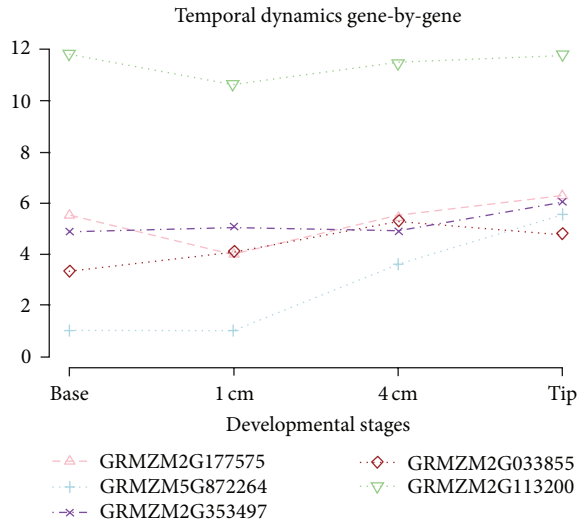


Figure 6: Gene expression curves for five significant TDE genes in maize leaf data. Mean expression curves are presented from two replicates during four developmental stages comparing temporal patterns each other in identifying statistically significant trajectories.

analysis and multivariate approaches, respectively, were further explored to reveal temporal relationships underlying biological processes based on gene ontology and functional network/pathway analysis. Sheep gene symbols with 21,865 genes were converted into human gene symbols with 15,343 using BioMart in R package [61]. In gene ontology analysis through Avadis NGS [62], 528 specific genes which were detected by uniquely HMM were further analyzed. Interestingly, as 63 females were sampled in the data, some of gender-specific GO terminologies among significant ontology terms were identified, that is, granulosa cell development and maternal placenta development as well as intermediate mesoderm formation, regulation of cell growth involved in cardiac muscle cell development, positive regulation of striated muscle contraction, response to stimulus involved in regulation of muscle adaptation, intermediate mesoderm formation, voluntary musculoskeletal movement, growth plate cartilage development, extracellular matrix, and so forth. The HMM-specific temporal dynamic gene sets were further investigated for coexpressed gene sets and functional network modules through ebdbNet and GeneNet in R package and GeneMANIA [63–65] as shown in Figures 5 and 6.

3.6. Simulation Studies. We show that dynamic methods outperform approaches that do not explicitly address the time series nature of the data in simulation studies for validation and evaluation. We evaluated the performance of dynamic methods with simulation studies in which temporal features are already known as gold standard TDE (temporally differentially expressed) gene lists. Gold standard gene lists contain entire information to mimic RNA-seq time series profile if a gene is differentially expressed (DE) or equally expressed (EE) over time as reference set to compare to the results obtained from both dynamic and static methods in terms of recall and precision measurement. To this end, we generated simulated RNA-seq datasets with expression profiling data points representing nondifferentially expressed and differentially expressed genes in a series of time points by using different values of autocorrelation parameter (φ). We generated data for equally expressed genes by sampling time series process parameters (w) of a gene in invertible Gaussian ARIMA process with $\varphi = 0$. We generated data for differentially expressed genes across time points in the same procedures as $\varphi = 0.1, 0.25, 0.5, 0.75$, and 0.9, respectively. After time series process, regression effects and autocorrelation parameters were simulated for 1000 genes, 4 simulated datasets were generated by setting the varying number of time points and replicates in a time point, $nT = 5$ and 10, $nR = 3$, and 5 to compute P value, FDR, and credible interval of each gene for static and dynamic methods and compared to gold standards to obtain true discovery rates in our simulated datasets.

4. Conclusion and Discussion

We first performed pairwise comparisons using two simple static methods, Audic-Claverie statistics and Fisher's exact test. The congruent set of both of them is highly overlapped and we reported the results of Fisher's exact test as more common method in Figure 1. The dataset came from a sheep model with two different healing systems at four different days. This dataset provides an excellent design for identification of temporally and simultaneously differentially expressed (TDE) genes as we have two conditions at each time. This type of time course is referred to as factorial time course experimental design. The authors of [53] took a union set of all these combinations of pairwise comparisons in condition and time point to identify TDE genes. These approaches might provide insights and intuitively simple static methods are alternative in small experiments in general. Evidently, the methods for time series dynamics are still in their infancy. However, those algorithms all do not consider dependency between samples in time course and they assume that all samples are independently distributed, though sequential correlation is obviously observed in data as shown in Figure 2. We noticed that basically patterns of detection of temporal changes by static and dynamic method are different, albeit they agree in some degree. That is, most of temporal genes at low and moderate expression levels are detected as significant genes in dynamic methods, whereas, due to power issues of parametric static pooling methods and simplification of pairwise methods, static methods do a good job at high expression levels. To confirm robustness and reliability of gene detection methods in time series, a comprehensive comparison and evaluation with varying parameter settings closer to RNA-seq real world is further needed. At low and moderate levels, many genes which were not detected by static methods but dynamic methods still showed log2-scaled FC ~4 up to 5. We sought to test the ability of dynamic methods whether or not those identified dynamic-unique TDE genes are genuinely differential expression or just by a random chance because expressions have been more affected by noise at low and moderate

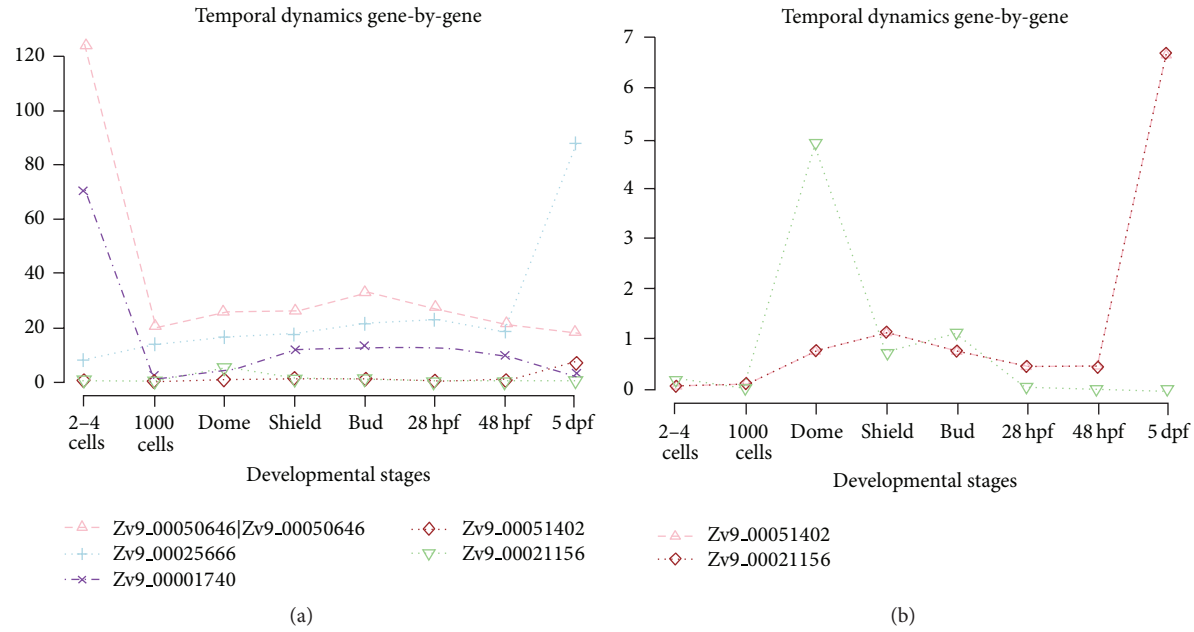


FIGURE 7: (a) Gene expression curves for five TDE genes in zebrafish data. Expression curves are presented during eight developmental stages comparing temporal patterns to each other in identifying statistical significant trajectories. (b) Two specific genes in gene-by-gene temporal dynamics with low expression levels via SETI in zebrafish data from (a). SETI enables identification of significant temporal patterns at low expression levels which are not detected by other existing static methods.

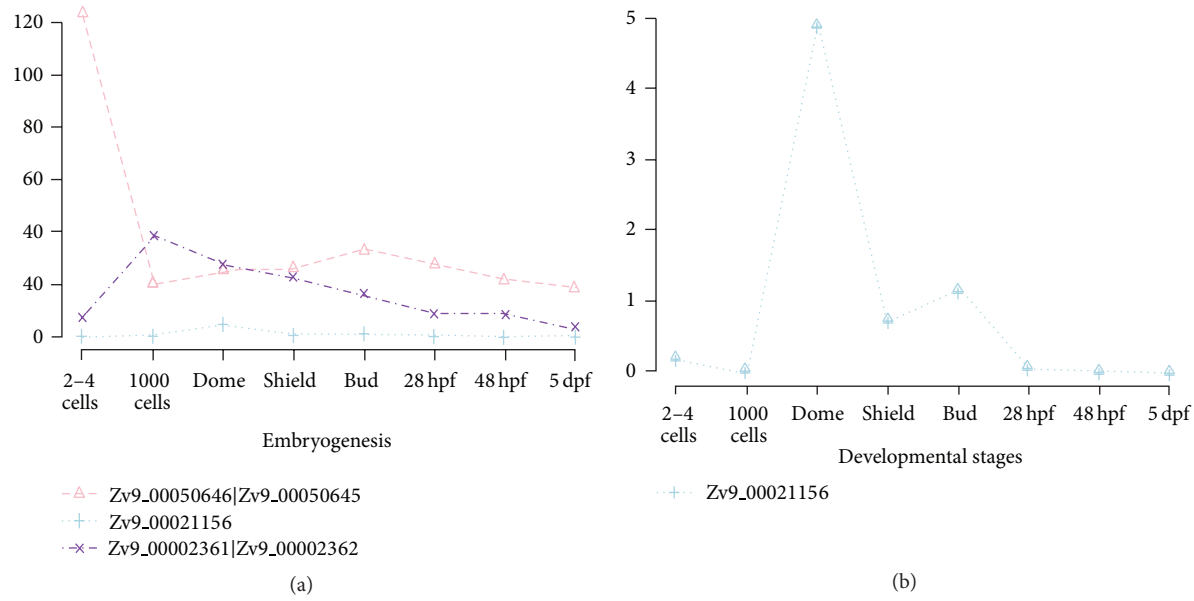


FIGURE 8: (a) A set of top cooperative TDE genes detected by Glass-d-score with FDR 0.05 as a candidate of cooperative gene pairs in coexpression in zebrafish data. Glass-d-score and corresponding FDR at cutoff 0.05 by resampling procedure under the null hypothesis that there are no TDE patterns across samples and those are shuffled with 1,000 repetitions. (b) One specific gene in coupled temporal dynamics with low expression level via SETI and Glass-d-score from (a). Glass-d-score is robust in identifying coexpressed genes over time at low expression levels.

levels in microarray, even though RNA-seq quality when compared to microarray has been much improved for now. In Figure 4, to assess temporally differentially expressed genes, we incorporated SETI with HMM algorithm in a sheep model within each condition to see a variety of time-varying trajectories since HMM provides only patterns of hidden latent variables (DE/EE) at developmental stages. Left panel shows three candidate genes at low and moderate levels and right panel shows another three candidate genes at high expression. To examine biological meanings in TDE

genes detected by only HMM, we further performed gene clustering coexpression patterns to see if those gene sets have possibility of false negatives in altered expression of cooperative genes and gene functional pathway analyses. Notably, we confirmed meaningful biological functionalities and temporal patterns in dynamic specific TDE genes in downstream analyses, gene clustering, gene ontology, and pathway/network analysis. Interestingly, as 63 females were sampled in the data, some of gender-specific GO terminologies among significant ontology terms were identified, that is, granulosa cell development and maternal placenta development, intermediate mesoderm formation, regulation of cell growth involved in cardiac muscle cell development, positive regulation of striated muscle contraction, response to stimulus involved in regulation of muscle adaptation, intermediate mesoderm formation, voluntary musculoskeletal movement, growth plate cartilage development, extracellular matrix, and so forth. Consistently, HMM, SETI, and AR(1) model that account for time dependency Markovian property in the models identified more of statistically significant TDE genes than static methods regardless of expression levels. In summary, the approaches we described use a developed unified dynamic test framework that includes SETI with statistical significance testing, ranking temporal genes by AR(1) modeling and posterior probability of autocorrelation parameter, and HMM to classify temporal dynamic patterns. These methods seem to be robust regardless of the magnitude of expression (see Figures 7(b) and 8(b), and Supplementary Tables) and more sensitive than static methods as shown in Supplementary TDE Tables; moreover, TDE genes detected by dynamic specific methods were confirmed as temporal dynamics in clustering patterns and biologically significant modules in network analysis implying that the gene sets were not identified as false negative genes in static methods that samples over time are assumed independently. We anticipate that temporal RNA-seq experiments will be widely performed in the near future due to reduced sequencing cost and the rich information carried by these experiments. In this paper, we consider several statistical approaches that can explicitly model the time series nature in the data. We discussed the limitations of simple static pairwise comparison methods for time series data analysis; dynamic statistical framework for RNA-seq read count with statistical evolutionary trajectory index measure; autoregressive time-lagged AR(1) model; hidden Markov model; pairwise and multiple comparisons among trajectories to investigate coupled bivariate dependency between distinct SETIs; and pathway/network analysis in transcriptome data based on detected temporally differentially expressed genes. Thus, this study covers critical issues that have not been systematically addressed in temporal RNA-seq data and we hope this will motivate more rigorous developments of novel methods to model and analyze RNA-seq data. Of particular interest will be the extension of these methods to combined time series datasets from RNA-seq, proteomics, and metabolomics for *in silico* cell/organism predictive modeling [6]. In addition, it will facilitate cross-species comparative analyses of temporal gene expression to investigate developmental processes and disease progression such as aging and virus-mediated immune disease dynamics. Deep sequencing of mRNAs has been a popular and effective approach for quantification of alternative splicing events, and it is well known that more than 90 percent of human genes have multiple isoforms to produce different protein structures. Thus, an important future direction is also to extend the statistical framework of our dynamic methods to incorporate the characterization of isoform diversity in time course in detecting differential expression.

Acknowledgments

This work was supported in part by NIH GM094780 (J. P. Noonan), GM59507 (H. Zhao) and NSF DMS 1106738 (H. Zhao).

References

[1] S. Marguerat and J. Bähler, "RNA-seq: from technology to biology," *Cellular and Molecular Life Sciences*, vol. 67, no. 4, pp. 569–579, 2010.

[2] B. A. Friedman and T. Maniatis, "ExpressionPlot: a web-based framework for analysis of RNA-Seq and microarray gene expression data," *Genome Biology*, vol. 12, p. R69, 2011.

[3] D. Ghosh and Z. S. Qin, "Statistical issues in the analysis of ChIP-seq and RNA-seq data," *Genes*, vol. 1, no. 2, pp. 317–334, 2010.

[4] L. M. McIntyre, K. K. Lopiano, A. M. Morse et al., "RNA-seq: technical variability and sampling," *BMC Genomics*, vol. 12, p. 293, 2011.

[5] Z. Fang and X. Cui, "Design and validation issues in RNA-seq experiments," *Briefings in Bioinformatics*, vol. 12, no. 3, Article ID bbr004, pp. 280–287, 2011.

[6] J. R. Karr, J. C. Sanghvi, D. N. Macklin et al., "A whole cell computational model predicts phenotype from genotype," *Cell*, vol. 150, no. 2, pp. 389–401, 2012.

[7] S. M. Wang, "Understanding SAGE data," *Trends in Genetics*, vol. 23, no. 1, pp. 42–50, 2007.

[8] A. Schliep, A. Schönhuth, and C. Steinhoff, "Using hidden Markov models to analyze gene expression time course data," *Bioinformatics*, vol. 19, no. 1, pp. i255–i263, 2003.

[9] R. Gottardo, A. E. Raftery, K. Yee Yeung, and R. E. Bumgarner, "Bayesian robust inference for differential gene expression in microarrays with multiple samples," *Biometrics*, vol. 62, no. 1, pp. 10–18, 2006.

[10] X. Han, W. K. Sung, and L. Feng, "Pem: a general statistical approach for identifying differentially expressed genes in time-course cDNA microarray experiment without replicate," in *Proceedings of the IEEE Computational Systems Bioinformatics Conference*, pp. 123–132, 2006.

[11] Y. C. Tai and T. P. Speed, "A multivariate empirical Bayes statistic for replicated microarray time course data," *Annals of Statistics*, vol. 34, no. 5, pp. 2387–2412, 2006.

[12] Y. C. Tai and T. P. Speed, "On gene ranking using replicated microarray time course data," *Biometrics*, vol. 65, no. 1, pp. 40–51, 2009.

[13] P. Ma, W. Zhong, and J. S. Liu, "Identifying differentially expressed genes in time course microarray data," *Statistics in Biosciences*, vol. 1, no. 2, pp. 144–159, 2009.

[14] Y. Goltsev and D. Papatsenko, "Time warping of evolutionary distant temporal gene expression data based on noise suppression," *BMC Bioinformatics*, vol. 10, p. 353, 2009.

[15] M. J. Aryee, J. A. Gutiérrez-Pabello, I. Kramnik, T. Maiti, and J. Quackenbush, "An improved empirical bayes approach to estimating differential gene expression in microarray time-course data: BETR (Bayesian Estimation of Temporal Regulation)," *BMC Bioinformatics*, vol. 10, p. 409, 2009.

[16] J. D. Storey, W. Xiao, J. T. Leek, R. G. Tompkins, and R. W. Davis, "Significance analysis of time course microarray experiments," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 36, pp. 12837–12842, 2005.

[17] T. Park, S. G. Yi, S. Lee et al., "Statistical tests for identifying differentially expressed genes in time-course microarray experiments," *Bioinformatics*, vol. 19, no. 6, pp. 694–703, 2003.

[18] W. Xu, J. Seok, M. N. Mindrinos et al., "Human transcriptome array for high-throughput clinical studies," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 9, pp. 3707–3712, 2011.

[19] Y. Liang, B. Tayo, X. Cai, and A. Kelemen, "Differential and trajectory methods for time course gene expression data," *Bioinformatics*, vol. 21, no. 13, pp. 3009–3016, 2005.

[20] A. A. Kalaitzis and N. D. Lawrence, "A simple approach to ranking differentially expressed gene expression time courses through Gaussian process regression," *BMC Bioinformatics*, vol. 12, p. 180, 2011.

[21] J. J. Song, H. J. Lee, J. S. Morris, and S. Kang, "Clustering of time-course gene expression data using functional data analysis," *Computational Biology and Chemistry*, vol. 31, no. 4, pp. 265–274, 2007.

[22] Y. Luan and H. Li, "Clustering of time-course gene expression data using a mixed-effects model with B-splines," *Bioinformatics*, vol. 19, no. 4, pp. 474–482, 2003.

[23] I. Sohn, K. Owzar, S. L. George, S. Kim, and S. H. Jung, "A permutation-based multiple testing method for time-course microarray experiments," *BMC Bioinformatics*, vol. 10, p. 336, 2009.

[24] S. C. Billups, M. C. Neville, M. Rudolph, W. Porter, and P. Schedin, "Identifying significant temporal variation in time course microarray data without replicates," *BMC Bioinformatics*, vol. 10, p. 96, 2009.

[25] J. T. Leek, E. Monsen, A. R. Dabney, and J. D. Storey, "EDGE: extraction and analysis of differential gene expression," *Bioinformatics*, vol. 22, no. 4, pp. 507–508, 2006.

[26] G. A. Churchill, "Fundamentals of experimental design for cDNA microarrays," *Nature Genetics*, vol. 32, no. 5, pp. 490–495, 2002.

[27] J. B. Fan, M. S. Chee, and K. L. Gunderson, "Highly parallel genomic assays," *Nature Reviews Genetics*, vol. 7, no. 8, pp. 632–644, 2006.

[28] J. Ernst and Z. Bar-Joseph, "STEM: a tool for the analysis of short time series gene expression data," *BMC Bioinformatics*, vol. 7, p. 191, 2006.

[29] Y. Yuan, Y. P. Chen, S. Ni et al., "Development and application of a modified dynamic time warping algorithm (DTW-S) to analyses of primate brain expression time series," *BMC Bioinformatics*, vol. 12, p. 347, 2011.

[30] Z. Bar-Joseph, G. Gerber, I. Simon, D. K. Gifford, and T. S. Jaakkola, "Comparing the continuous representation of time-series expression profiles to identify differentially expressed genes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 18, pp. 10146–10151, 2003.

[31] Y. Yuan, C. T. Li, and R. Wilson, "Partial mixture model for tight clustering of gene expression time-course," *BMC Bioinformatics*, vol. 9, p. 287, 2008.

[32] G. C. Tseng and W. H. Wong, "Tight clustering: a resampling-based approach for identifying stable and tight patterns in data," *Biometrics*, vol. 61, no. 1, pp. 10–16, 2005.

[33] S. Dudoit, Y. H. Yang, M. J. Callow, and T. P. Speed, "Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments," *Statistica Sinica*, vol. 12, no. 1, pp. 111–139, 2002.

[34] T. T. Nguyen, R. R. Almon, D. C. DuBois, W. J. Jusko, and I. P. Androulakis, "Importance of replication in analyzing time-series gene expression data: corticosteroid dynamics and circadian patterns in rat liver," *BMC Bioinformatics*, vol. 11, p. 297, 2010.

[35] M. J. Nueda, J. Carbonell, I. Medina, J. Dopazo, and A. Conesa, "Serial expression analysis: a web tool for the analysis of serial gene expression data," *Nucleic Acids Research*, vol. 38, no. 2, Article ID gkq488, pp. W239–W245, 2010.

[36] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics*, vol. 26, no. 1, pp. 139–140, 2010.

[37] S. Anders and W. Huber, "Differential expression analysis for sequence count data," *Genome Biology*, vol. 11, no. 10, p. R106, 2010.

[38] G. K. Smyth, "Linear models and empirical bayes methods for assessing differential expression in microarray experiments," *Statistical Applications in Genetics and Molecular Biology*, vol. 3, no. 1, 2004.

[39] J. H. Bullard, E. Purdom, K. D. Hansen, and S. Dudoit, "Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments," *BMC Bioinformatics*, vol. 11, p. 94, 2010.

[40] J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad, "RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays," *Genome Research*, vol. 18, no. 9, pp. 1509–1517, 2008.

[41] S. Tarazona, F. Garcia-Alcalde, J. Dopazo, A. Ferre, and A. Conesa, "Differential expression in RNA-Seq: a matter of depth," *Genome Research*, vol. 21, pp. 2213–2223, 2011.

[42] T. J. Hardcastle and K. A. Kelly, "BaySeq: empirical Bayesian methods for identifying differential expression in sequence count data," *BMC Bioinformatics*, vol. 11, p. 422, 2010.

[43] P. L. Auer and R. W. Doerge, "Statistical design and analysis of RNA sequencing data," *Genetics*, vol. 185, no. 2, pp. 405–416, 2010.

[44] P. L. Auer, S. Srivastava, and R. W. Doerge, "Differential expression—the next generation and beyond," *Briefings in Functional Genomics*, 2011.

[45] J. Lee, Y. Ji, S. Liang, G. Cai, and P. Muller, "On differential gene expression using RNA-Seq data," *Cancers Information*, vol. 10, pp. 205–215, 2011.

[46] A. Oshlack, M. D. Robinson, and M. D. Young, "From RNA-seq reads to differential expression results," *Genome Biology*, vol. 11, no. 12, p. 220, 2010.

[47] C. Trapnell, B. A. Williams, G. Pertea et al., "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation," *Nature Biotechnology*, vol. 28, no. 5, pp. 511–515, 2010.

[48] M. Yuan and C. Kendziorski, "Hidden Markov models for microarray time course data in multiple biological conditions," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1323–1332, 2006.

[49] M. Jäger, C. E. Ott, J. Grünhagen et al., "Composite transcriptome assembly of RNA-seq data in a sheep model for delayed bone healing," *BMC Genomics*, vol. 12, p. 158, 2011.

[50] P. Li, L. Ponnala, N. Gandotra et al., "The developmental dynamics of the maize leaf transcriptome," *Nature Genetics*, vol. 42, no. 12, pp. 1060–1067, 2010.

[51] A. Pauli, E. Valen, M. F. Lin et al., "Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis," *Genome Research*, vol. 22, pp. 577–591, 2012.

[52] J. Lienau, K. Schmidt-Bleek, A. Peters et al., "Insight into the molecular pathophysiology of delayed bone healing in a sheep model," *Tissue Engineering A*, vol. 16, no. 1, pp. 191–199, 2010.

[53] H. Schell, M. S. Thompson, H. J. Bail et al., "Mechanical induction of critically delayed bone healing in sheep: radiological and biomechanical results," *Journal of Biomechanics*, vol. 41, no. 14, pp. 3066–3072, 2008.

[54] J. L. Hay and A. N. Pettitt, "Bayesina analysis of a time series of counts with covariates: an application to the control of an infectious disease," *Biostatistics*, vol. 2, no. 4, pp. 433–444, 2011.

[55] M. L. T. Lee, F. C. Kuo, G. A. Whitmore, and J. Sklar, "Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, no. 18, pp. 9834–9839, 2000.

[56] C. Spyrou, R. Stark, A. G. Lynch, and S. Tavaré, "BayesPeak: Bayesian analysis of chip-seq data," *BMC Bioinformatics*, vol. 10, p. 299, 2009.

[57] M. A. Newton, C. M. Kendziorski et al., *Parametric Empirical Bayes Methods for Microarrays in the Analysis of Gene Expression Data: Methods and Software*, Springer, New York, NY, USA, 2003.

[58] N. Wiener, "The theory of prediction," in *Modern Mathematics for Engineers*, E. Beckenbach, Ed., McGraw-Hill, New York, NY, USA, 1956.

[59] S. Audic and J. M. Claverie, "The significance of digital gene expression profiles," *Genome Research*, vol. 7, no. 10, pp. 986–995, 1997.

[60] G. Glass, "Integrating findings: the meta-analysis of research," *Review of Research in Education*, vol. 5, pp. 351–379, 1977.

[61] "bioMart in bioconductor," http://www.bioconductor.org/.

[62] Strand Life Sciences Pvt. Ltd., *Avadis NGS Version: Version 1.3.0*, Strand Scientific Intelligence, San Francisco, NC, USA, 2012.

[63] D. Warde-Farley, S. L. Donaldson, O. Comes et al., "The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function," *Nucleic Acids Research*, vol. 38, no. 2, Article ID gkq537, pp. W214–W220, 2010.

[64] "GeneNet in bioconductor," http://www.bioconductor.org/.

[65] "ebdbNet in bioconductor," http://www.bioconductor.org/.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 582358, 7 pages
http://dx.doi.org/10.1155/2013/582358

Research Article

Gene Entropy-Fractal Dimension Informatics with Application to Mouse-Human Translational Medicine

T. Holden, E. Cheung, S. Dehipawala, J. Ye, G. Tremberger Jr., D. Lieberman, and T. Cheung

Queensborough Community College of CUNY, 222-05 56th Avenve Bayside, NY 11364, USA

Correspondence should be addressed to T. Holden; tholden@qcc.cuny.edu

Received 6 October 2012; Accepted 5 February 2013

Academic Editor: Tun-Wen Pai

DNA informatics represented by Shannon entropy and fractal dimension have been used to form 2D maps of related genes in various mammals. The distance between points on these maps for corresponding mRNA sequences in different species is used to study evolution. By quantifying the similarity of genes between species, this distance might be indicated when studies on one species (mouse) would tend to be valid in the other (human). The hypothesis that a small distance from mouse to human could facilitate mouse to human translational medicine success is supported by the studied ESR-1, LMNA, Myc, and RNF4 sequences. ID1 and PLCZ1 have larger separation. The collinearity of displacement vectors is further analyzed with a regression model, and the ID1 result suggests a mouse-chimp-human translational medicine approach. Further inference was found in the tumor suppression gene, p53, with a new hypothesis of including the bovine PKM2 pathways for targeting the glycolysis preference in many types of cancerous cells, consistent with quantum metabolism models. The distance between mRNA and protein coding CDS is proposed as a measure of the pressure associated with noncoding processes. The Y-chromosome DYS14 in fetal micro chimerism that could offer protection from Alzheimer's disease is given as an example.

1. Introduction

When a nucleotide in a DNA sequence is different from the preceding nucleotide, this is defined as a nucleotide fluctuation. The nucleotide fluctuations of a DNA sequence can be studied as a series using the nucleotide atomic number of the nucleotide A, T, C, and G. A recent study on such fluctuation in the FOXP2 gene has been reported [1]. The fractal dimension and Shannon entropy was found to have a negative correlation ($R^2 \sim 0.85$ $N = 12$) for the FOXP2 regulated accelerated conserved noncoding sequences in human fetal brain. This paper uses a 2D mapping of the Shannon entropy and fractal dimension to determine displacement vectors, which could serve as a marker for the evolutionary differences between mouse and human DNA in clinically important gene sequences. The hypothesis that displacement vectors having small separation would facilitate the mouse to human translational medicine success would be testable with gene therapy cases. The selected gene candidates in this report are based on new discoveries reported in and around September 2012. The ESR1 neuronal estrogen receptor was reported by Rockefeller University to be a single "mommy" gene such that malfunction deletion would suppress motherhood behavior [2]. Successful control of Hutchinson-Gilford progeria syndrome in children by correcting the mutated LMNA lamin A protein was reported by Harvard Medical School [3]. The Myc myelocytomatosis oncogene was reported by US National Institutes of Health to be a universal amplifier for cancer already turned on by another process [4]. The RNF4, RING finger protein 4 with zinc finger motif, was reported by UK Dundee University to be necessary for human response to DNA damage [5]. The ID1, a DNA-binding protein inhibitor, associated with aggressive nonstandard breast cancer cells could be controlled by cannabidiol in cannabis [6]. The PLCZ1, phospholipase C Zeta 1, was reported to be delivered by the sperm to control egg activation [7]. Calibration based on 16S rRNA (human and mouse) enables a relative measure of the evolutionary pressure of the above genes between human and mouse. The HAR1 sequence with 118-bp, is the fastest evolving human sequence as compared to the chimp. It contains 18

point substitutions occurring over a span of 5 million years when comparing the human to the chimpanzee. However, the same 118-bp region only contains two-point substitutions over a span of 300 million years when comparing the chimpanzee to the chicken [8]. The inclusion of HAR1 in the calibration should set an upper limit for the displacement vector magnitude.

2. Materials and Methods

The data used in this study was downloaded from Genbank and the accession information is listed [9–18]. The HAR1 human and chimp sequences were downloaded with information from [8].

A sequence with a relatively low nucleotide variety would have low Shannon entropy (more constraint) in terms of the set of 16 possible dinucleotide pairs. A sequence's entropy can be computed as the sum of $(p_i) * \log(p_i)$ over all states i, and the probability p_i can be obtained from the empirical histogram of the 16 di-nucleotide-pairs. The maximum entropy is 4 binary bits per pair for 16 possibilities (2^4). For mononucleotide consideration, the maximum entropy is two bits per mono with four possibilities (2^2). The mononucleotide entropy is correlated to dinucleotide entropy $R^2 > 0.9$ for all studied sequences in the project.

Fractal dimension analysis on data series can be used in the study of correlated randomness. Among the various fractal dimension methods, the Higuchi fractal method is well suited for studying fluctuation [19]. The spatial intensity (Int) series with equal intervals is be used to generate a difference series $(\text{Int}(j) - \text{Int}(i))$ for different lags $(j - i)$ in the spatial variable. The nonnormalized apparent length of the spatial series curve is simply $L(k) = \Sigma|\text{Int}(j) - \text{Int}(i)|$ for all $(j - i)$ pairs that equal to k. The number of terms in a k-series varies, and normalization must be used to get the series length. If the $\text{Int}(i)$ is a fractal function, then the $\log(L(k))$ versus $\log(1/k)$ should be a straight line with the slope equal to the fractal dimension. Higuchi incorporated a calibration division step such that the maximum theoretical value is calibrated to the topological value of 2. The detailed calculation is given in the literature [19]. The Higuchi fractal algorithm used in this project was calibrated with the Weierstrass function. This function has the form $W(x) = \Sigma\, a^{-nh} \cos(2\pi a^n x)$ for $n = 0, 1, 2, 3, \ldots$. The fractal dimension of the Weierstrass function is given by $(2 - h)$, where h takes on an arbitrary value between zero and one.

Although the Higuchi method was originally developed for time series data, Fractal dimension analysis is an established method to analyze DNA sequences and other finite progressions [20]. By comparing the fractal dimension for a concatenated infinite sequence of known fractal dimension, we obtain results similar to those shown in Figure 8 of [21]. For the lengths of sequences analyzed in this paper, the error is about 1% or less, corresponding to about one fifth of the variation in fractal dimension seen in this paper. Thus, we conclude that the current analysis is justified for these sequences.

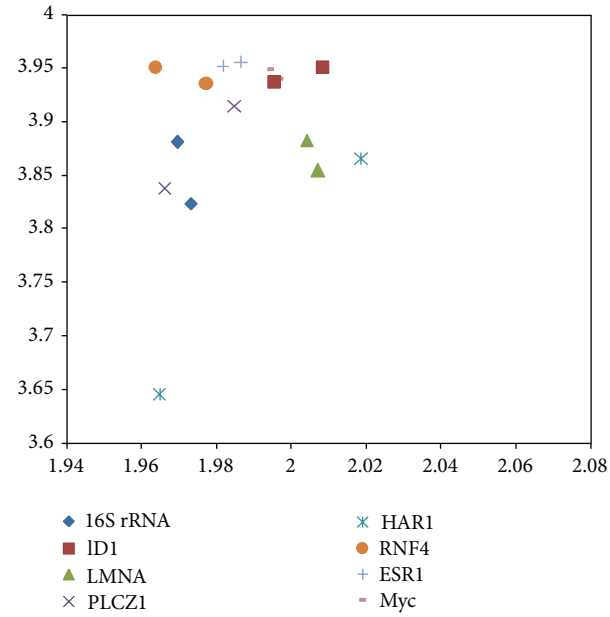


Figure 1: The mRNA 2D map of the studied mouse-human pairs. The y-axis represents dinucleotide entropy in bits per symbol, and x-axis presents the fractal dimension. 16S rRNA (diamond), ID1 (square), PLCZ1 (cross), RNF4 (circle), ESR1 (plus), and Myc (bar) have lower fractal dimensions for human. The LMNA (triangle) and HAR1 have higher fractal dimension for human.

3. Results of Fractal Analysis

The mRNA and protein coding CDS 2D maps of entropy and fractal dimension of the studied mouse-human pairs are shown below in Figures 1 and 2, respectively. The mRNA human sequences except LMNA and HAR1 show lower fractal dimension as compared to the mouse counterparts. The CDS human sequences except LMNA, HAR1, and RNF4 show lower fractal dimension as compared to the mouse counterparts. Furthermore, the separation from one point to another could be represented by a displacement vector. A regression model is applicable for ID1 human variant 1, human variant 2, and chimp given the collinearity of the displacement vectors. The ID1 regression result is displayed in Figure 3. The graph scale is identical to that of Figures 1 and 2 for easy comparison. The x-axis fractal dimension should not be interpreted as the independent variable.

4. Discussion

The mouse to human difference is represented by the coordinate separation in Figure 1 (mRNA sequences) and Figure 2 (CDS sequences). HAR1 has the most separation in terms of coordinates in Figure 1, consistent with the labeling of the most accelerated region, given 18 point mutation from chimp to human in 118-bp. The HAR1 mouse counterpart is close to HAR1 chimp counterpart and has a fractal dimension of 1.945 and 3.657 bits per symbol (not displayed). The CDS map in Figure 2 shows ID1 having the most separation, followed by PLCZ1. BLAST comparison of mouse versus human results

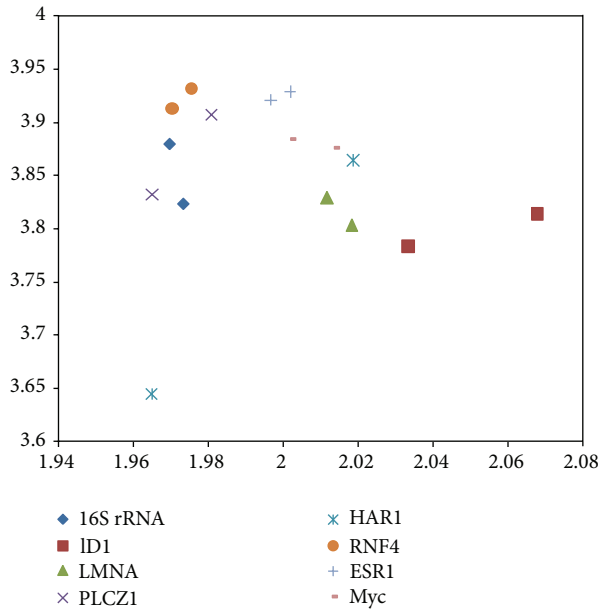


Figure 2: The protein coding CDS 2D map of the studied mouse-human pairs. The y-axis represents di-nucleotide entropy in bits per symbol, and x-axis presents the fractal dimension. 16S rRNA (diamond), ID1 (square), PLCZ1 (cross), ESR1 (plus), and Myc (bar), have lower fractal dimension for human. LMNA (triangle), HAR1 (star), and RNF4 (circle) have higher fractal dimension for human.

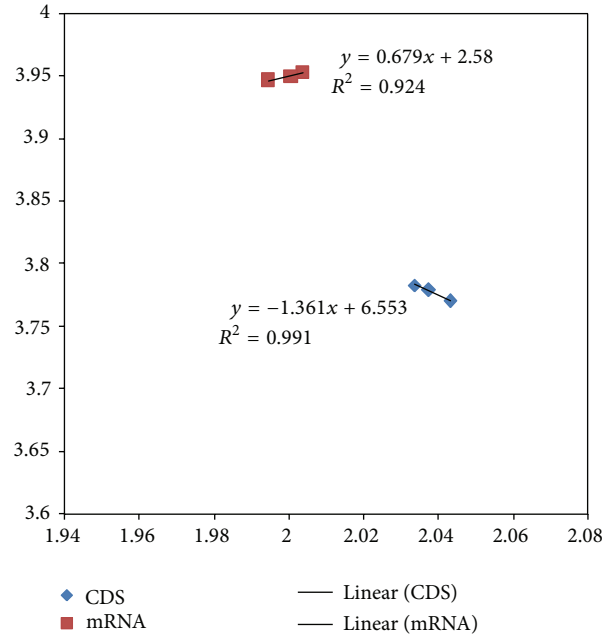


Figure 3: The regression model of human ID1 variant1, human ID1 variant2, and chimp ID1. The y-axis represents di-nucleotide entropy in bits per symbol, and x-axis presents the fractal dimension. The graph scale is kept identical to that of Figures 1 and 2 for easy comparison. The CDS sequence (diamond) regression has R^2 of 0.991 and an adjusted R^2 of 0.983 (the chimp has the highest fractal dimension). The mRNA (square) sequence regression has R^2 of 0.924 (the chimp is in the middle among the three data points).

show E-value of zero for PLCZ1, suggesting that the entropy-fractal dimension 2D map can have a finer resolution. A large coordinate separation would be expected to represent very different sets of regulatory pathways from mouse to human. When comparing Figure 1 with Figure 2, the spreading of CDS data points as compared to the mRNA data points is dominated by ID1 coordinate change. For example, the coordinate change of CDS-ID1 from mouse to human would be comparable to the HAR1 separation representing an evolutionary aspect from chimp to human. Furthermore, as collinearity in displacement vectors could be represented by regression, the result of the coordinate changes in the CDS map of Figure 2 from that the mRNA map of Figure 1 increases the collinearity of the displacement vectors. For example, for ID1 in human variant 1, human variant 2, and chimp, the coordinate changes from mRNA to CDS have resulted in an increasing R^2 from 0.93 (mRNA) to 0.99 (CDS) as displayed in Figure 3.

If one defines evolutionary pressure as the cause of species transformation, then CDS pressure could be defined as the cause of informatics transformation from mRNA to CDS and, correspondingly, mRNA pressure be defined as the cause of informatics transformation from gene to mRNA. A displacement vector in Figure 4 (denoted by a line) would represent the mRNA pressure in ID1 for human, and mouse also. A displacement vector in the 2D map formed in comparing Figures 1 and 2 would represent the CDS pressure. The collinearity of displacement vectors modeled as regression would represent the evolutionary pressure from chimp to human. A vector carries two pieces of information. A displacement vector carries separation or distance or magnitude information and directionality information such as from mRNA to CDS and chimp to human. A displacement vector analysis of Y-chromosome DYS14 in fetal microchimerism was performed, and the result is displayed in Figure 5 where the selection of higher fractal dimension in mRNA pressure and CDS pressure is clearly demonstrated. The retention of DYS14 in a mother's brain was also reported to be consistent with protection for Alzheimer's disease for mothers who had sons [22].

A nucleotide sequence carries the informatics needed for a cell to live. A cell would continue to access the informatics throughout its lifetime. Average and standard deviation cannot represent the fluctuation or ordering of the nucleotides. Shannon entropy is a measure of the information content and fractal dimension could be interpreted as a measure of information order. In analogy to the Gas Law where pressure would be the cause of a temperature change given volume content, a displacement vector in the 2D map could be used as a marker for a pressure that would cause a fractal dimension change. Given the relatively large separation of ID1 as compared to the other studied sequences in Figure 2, a mouse-chimp-human approach would have supporting evidence. The data of other animals' ID1 sequences is shown in Figure 6, and using a mouse-monkey-human approach seems justified as well. Similarly, the Figure 7 CDS 2D map for the p-53 gene,

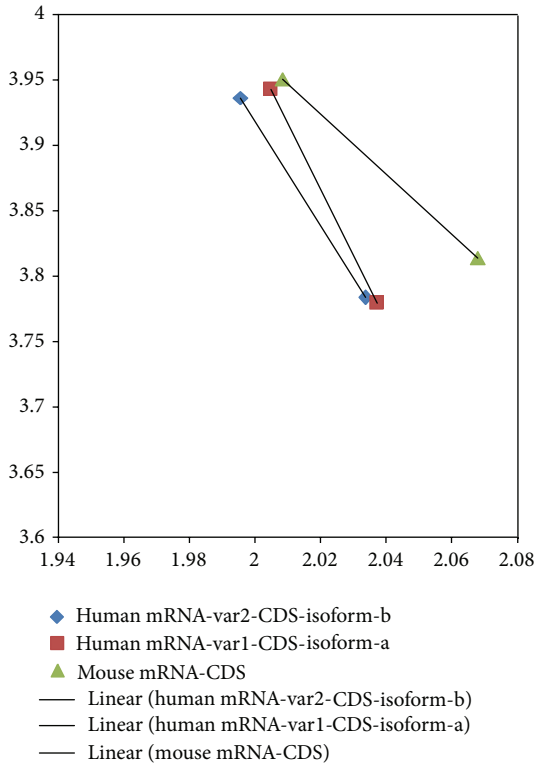


Figure 4: Displacement vector from mRNA to CDS for human ID1, and mouse ID1. The y-axis represents di-nucleotide entropy in bits per symbol and x-axis presents the fractal dimension. The separation or distance is shown as the length of the displayed line and the direction is from mRNA coordinates (upper diamond, upper square and upper triangle) to CDS coordinates (lower diamond, lower square and lower triangle).

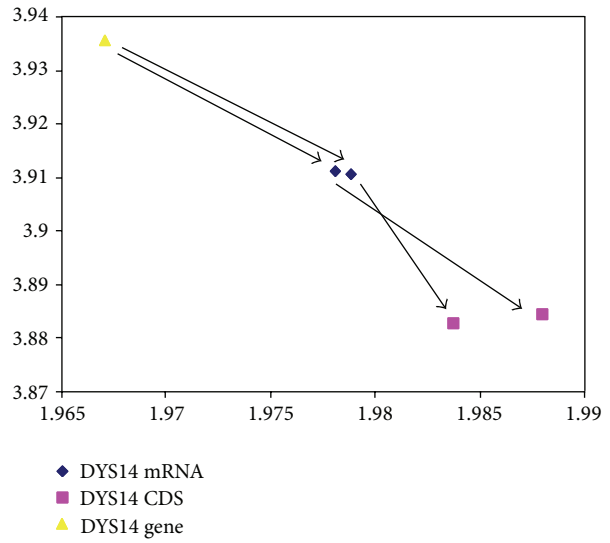


Figure 5: Entropy-fractal dimension map for Y-chromosome DYS14 Gene, mRNA, and CDS. The y-axis represents di-nucleotide entropy in bits per symbol, and x-axis presents the fractal dimension. The DYS14 gene (triangle) has the lowest fractal dimension, and the DYS14 CDS variant-1 and variant-2 (squares) are of higher fractal dimension, displayed as two data points in the lower right corner. DYS14 mRNA variant-1 and variant-2 (diamonds) have intermediate fractal dimension in comparison. The arrows represent the displacement vectors.

known for its role in tumor suppression [23], would suggest a mouse-dog-human approach also to be valid. The collinearity represented by a regression gives an R^2 of 0.96, with adjusted R^2~0.93 (Figure 7). Recent advance in quantum metabolism modeling provides supporting evidence of natural section pressure on glycolysis preference over oxidative phosphorylation in cancerous environment [24]. The discovery of PKM2 dimeric form in elevated levels in many cancers has echoed the Warburg Effect in oncology and explained the rapid glycolysis [25]. The PKM2 evolutionary paths can be visualized in an entropy-fractal dimension 2D map (Figure 8). Targeting the PKM2 pathways could be a possible cancer therapy in the standard human-mouse model. The human-bovine (Bos Taurus) hypothesis could be a supplemental approach, especially for those conditions with lower fractal dimension value sequences among the seven PKM2 variants in human. The entropy-fractal dimension 2D map is a very sensitive tool for comparative analysis. An analogy would be a Fabry-Perot interferometer for resolving wavelengths given that the interference order is already selected. Translational medicine based on genetics would benefit from the entropy-fractal dimension 2D map analysis in the selection of a species model.

Other fractal analysis results with the aim of translational medicine application have been reported. The H1N1 virus hemagglutinin (HA) sequences from various strains have been classified with correlation matrix fractal dimension values ranging from 2.29 to 2.32 in using a DNA representation via the Voss indicator function [20, 26]. The multifractal property of myeloma multiple TET2 mRNA Variant1 and Variant2 has been shown to converge to 1.26 in fractal dimension [27]. In fact, such DNA representation has been applied to generate DNA walk patterns with wavelet analysis that reveals hidden symmetries [28, 29]. On the broader chromosome level, it was reported that the chromosome-3 in Caenorhabditis elegans has coding regions averaging 1.306 and noncoding regions averaging 1.298 in fractal dimension values [30]. The fundamental computer science string representation for DNA sequences has also been studied. Assigning binary strings such that A = (00), T = (11), C = (01), and G = (10) have been used for the study of olfactory receptor OR1D2 sequences in human, chimp, and mouse [31]. Other popular DNA representation schemes can be found in a recent computer science review where the relative strengths of several assignment schemes were compared. For example, the Galois indicator sequence where A = 0, T = 2, C = 1, and G = 3 would work well in exon detection [32]. Regardless of the DNA representation scheme, the complexity of a sequence would be revealed by fractal analysis.

A new hypothesis that high fractal dimension sequences may be top level regulators (transcription factors) recently discussed in the ENCODE project would deserve further

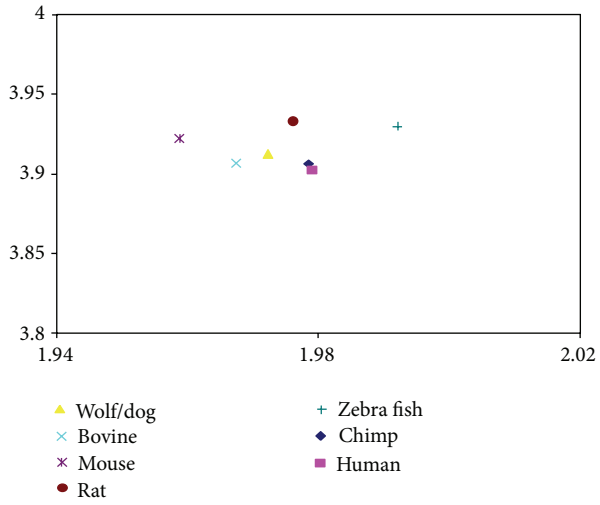


Figure 6: The protein coding CDS 2D map of the studied p53 sequences. The y-axis represents di-nucleotide entropy in bits per symbol, and x-axis presents the fractal dimension. The studied sequences included wolf/dog (triangle), bovine (cross), mouse (star), rat (circle), zebra fish (plus), chimp (diamond), and human (square).

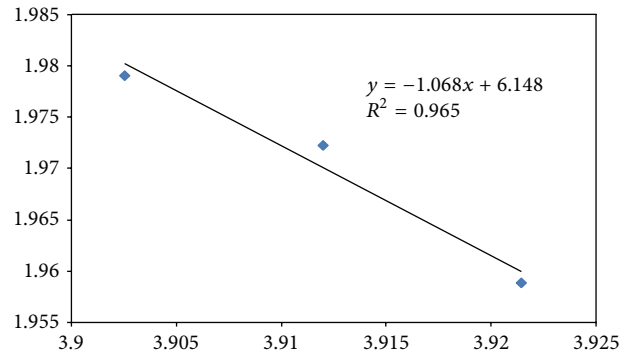


Figure 7: Entropy-fractal dimension for p53 CDS. The x-axis represents di-nucleotide entropy in bits per symbol, and y-axis presents the fractal dimension. Human has the highest fractal dimension, followed by wolf/dog, and mouse with the lowest fractal dimension.

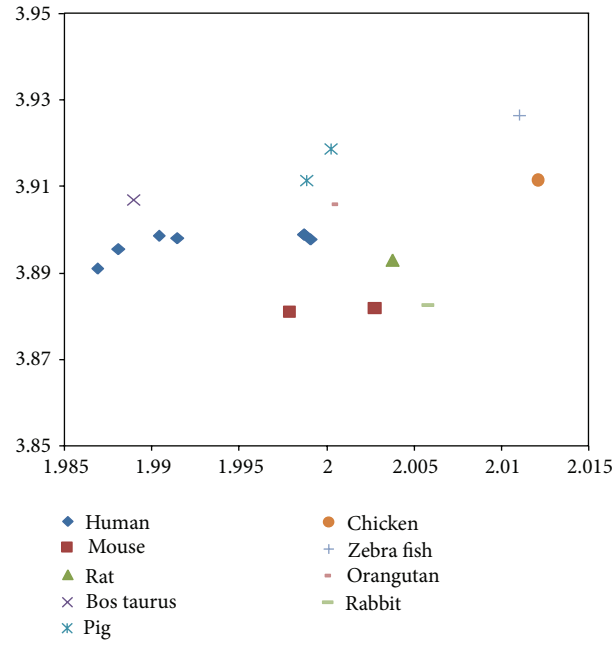


Figure 8: Entropy-fractal dimension for PKM2 CDS. The y-axis represents di-nucleotide entropy in bits per symbol, and x-axis presents the fractal dimension. The PKM2 of human (diamond, with 7 variants, gene no. 5315), mouse (square, with 2 variants, gene no. 18746), rat (triangle, gene no. 25630), bos Taurus (cross, gene no. 512571), pig (star, gene no. 100158154), chicken (circle, gene no. 396456), zebrafish (plus, gene no. 335817), orangutan (shorter bar, gene no. 100174114), and rabbit (longer bar, gene no. 100008676) are displayed.

investigation [33]. Other hypotheses, although not the main concern in translational medicine, could include high fractal dimension sequence as regulator for bioelectricity in microbes [34], optimal fractal dimension sequence for the photosynthesis genes involving quantum transport [35], and predicted entanglement process [36, 37].

5. Conclusions

The DNA gene sequence informatics represented by Shannon entropy and fractal dimension have been used to form 2D maps, and coordinate changes have been used in a displacement vector formulation for the studying of evolution with directionality. Although fractal dimension only mathematically applies to infinite fractal series, we found the error introduced by the finite size of our DNA sequences to be less than one fifth of the observed variation, thus justifying our analysis from a mathematical perspective. The hypothesis that small displacement vector from mouse to human could facilitate mouse to human translational medicine success has received support from the studied ESR-1, LMNA, Myc, and RNF4 in terms of their CDS and mRNA sequences. The collinearity of displacement vectors is further analyzed with a regression model, and the ID1 result suggests a mouse-chimp-human translational medicine approach. Other systems were studied with similar results, including the tumor suppression p53 within a mouse-wolf(dog)-human framework, leading to a new hypothesis of including the bovine PKM2 pathways for targeting the glycolysis preference in many types of cancerous cells, thus supplementing quantum metabolism studies as well. The displacement vector from mRNA coordinates to protein coding CDS coordinates could be a measure of the CDS pressure associated with non-coding process. The Y-chromosome DYS14 in fetal microchimerism is given as an example that CDS pressure, as well as mRNA pressure from gene to mRNA, would result in a higher fractal dimension sequence. A new hypothesis that high fractal dimension sequences could be top level transcription factors recently discussed in the ENCODE project deserves further investigation.

Acknowledgments

The project was partially supported by CUNY research grant (T. Holden). J. Ye thanks the NSF-REU program for student support. E. Cheung and S. Dehipawala thank QCC Physics Department for the hospitality. The authors thank the research groups cited in this paper for posting their data and software in the public domain.

References

[1] G. Tremberger Jr., S. Dehipawala, E. Cheung et al., "Fractal analysis of FOXP2 regulated accelerated conserved non-coding sequences in human fetal brain," *Engineering and Technology*, no. 67, pp. 881–886, 2012.

[2] A. C. Ribeiro, S. Musatov, A. Shteyler et al., "siRNA silencing of estrogen receptor-α expression specifically in medial preoptic area neurons abolishes maternal care in female mice," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 40, pp. 16324–16329, 2012.

[3] L. B. Gordon, M. E. Kleinmana, D. T. Miller et al., "Clinical trial of a farnesyltransferase inhibitor in children with Hutchinson-Gilford progeria syndrome," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 41, pp. 16666–16671, 2012.

[4] C. Y. Lin, J. Lovén, P. B. Rahl et al., "Transcriptional amplification in tumor cells with elevated c-Myc," *Cell*, vol. 151, no. 1, pp. 56–67, 2012.

[5] Y. Yin, A. Seifert, J. S. Chua, J.-F. Maure, F. Golebiowski, and R. T. Hay, "SUMO-targeted ubiquitin E3 ligase RNF4 is required for the response of human cells to DNA damage," *Genes & Development*, vol. 26, pp. 1196–1208, 2012.

[6] S. D. McAllister, R. Murase, R. T. Christian et al., "Pathways mediating the effects of cannabidiol on the reduction of breast cancer cell proliferation, invasion, and metastasis," *Breast Cancer Research and Treatment*, vol. 129, no. 1, pp. 37–47, 2011.

[7] M. Nomikos, K. Swann, and F. A. Lai, "Starting a new life: sperm PLC-zeta mobilizes the Ca^{2+} signal that induces egg activation and embryo development: an essential phospholipase C with implications for male infertility," *Bioessays*, vol. 34, pp. 126–134, 2012.

[8] K. S. Pollard, S. R. Salama, N. Lambert et al., "An RNA gene expressed during cortical development evolved rapidly in humans," *Nature*, vol. 443, no. 7108, pp. 167–172, 2006.

[9] "16S rRNA Human MT-RNR2 gene sequence," mouse gene/17725, http://www.ncbi.nlm.nih.gov/gene/4550.

[10] "ID1 Human gene sequence," mouse gene/15901, http://www.ncbi.nlm.nih.gov/gene/3397.

[11] "LMNA Human gene sequence," mouse gene/16905, http://www.ncbi.nlm.nih.gov/gene/4000.

[12] "PLCZ1 Human gene sequence," mouse gene/114875, http://www.ncbi.nlm.nih.gov/gene/89869.

[13] HAR1 Ref 8 Supplement Figure S2, pp. 44.

[14] "RNF4 Human gene sequence," mouse gene/19822, http://www.ncbi.nlm.nih.gov/gene/6047.

[15] "ESR1Human gene sequence," mouse gene/13982, http://www.ncbi.nlm.nih.gov/gene/2099.

[16] "Myc Human gene sequence," mouse gene/17869, http://www.ncbi.nlm.nih.gov/gene/4609.

[17] "DYS14 Human gene sequence," Approximately 35 copies of this gene are present in humans, but only a single, non-functional orthologous gene is found in mouse, http://www.ncbi.nlm.nih.gov/gene/7258.

[18] "p53 Human gene sequence," mouse gene/22059, wolf/dog gene/403869, zebra fish gene/30590, rat gene/24842, Pan troglodytes (chimp) gene/455214, bovine gene/281542, http://www.ncbi.nlm.nih.gov/gene/7157.

[19] T. Higuchi, "Approach to an irregular time series on the basis of the fractal theory," *Physica D*, vol. 31, no. 2, pp. 277–283, 1988.

[20] R. F. Voss, "Evolution of long-range fractal correlations and 1/f noise in DNA base sequences," *Physical Review Letters*, vol. 68, no. 25, pp. 3805–3808, 1992.

[21] P. Cristea, "An efficient algorithm for measuring fractal dimension of complex sequences," in *Proceedings of the Interdisciplinary Approaches in Fractal Analysis (IAFA '03)*, pp. 121–124, Bucharest, Romania, May 2003.

[22] W. F. N. Chan, C. Gurnot, T. J. Montine, J. A. Sonnen, K. A. Guthrie, and J. L. Nelson, "Male microchimerism in the human female brain," *PLoS One*, vol. 7, no. 9, Article ID e45592, 2012.

[23] A. G. Jegga, A. Inga, D. Menendez, B. J. Aronow, and M. A. Resnick, "Functional evolution of the p53 regulatory network through its target response elements," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 3, pp. 944–949, 2008.

[24] P. Davies, L. A. Demetrius, and J. A. Tuszynski, "Implications of quantum metabolism and natural selection for the origin of cancer cells and tumor progression," *AIP Advances*, vol. 2, Article ID 011101, 2012.

[25] H. R. Christofk, M. G. Vander Heiden, M. H. Harris et al., "The M2 splice isoform of pyruvate kinase is important for cancer metabolism and tumour growth," *Nature*, vol. 452, no. 7184, pp. 230–233, 2008.

[26] C. Cattani, "Fractals and hidden symmetries in DNA," *Mathematical Problems in Engineering*, vol. 2010, Article ID 507056, 31 pages, 2010.

[27] C. Cattani, G. Pierro, and G. Altieri, "Entropy and multifractality for the myeloma multiple TET 2 gene," *Mathematical Problems in Engineering*, vol. 2012, Article ID 193761, 14 pages, 2012.

[28] C. Cattani, "Complex representation of DNA sequences," *Communications in Computer and Information Science*, vol. 13, pp. 528–537, 2008.

[29] C. Cattani, "On the existence of wavelet symmetries in archaea DNA," *Computational and Mathematical Methods in Medicine*, vol. 2012, Article ID 673934, 21 pages, 2012.

[30] G. Pierro, "Sequence complexity of chromosome 3 in caenorhabditis elegans," *Advances in Bioinformatics*, vol. 2012, Article ID 287486, 12 pages, 2012.

[31] S. S. Hassan, P. P. Choudhury, B. S. Dayasagar, S. Chakraborty, R. Guha, and A. Goswami, "Understanding Genomic Evolutionof Olfactory Receptors through Fractal and Mathematical Morphology," *Nature Precedings*, 2011, http://precedings.nature.com/documents/5674/version/1.

[32] S. Arniker and H. Kwan, "Advanced numerical representation of DNA sequences," in *Proceedings of the International Conference on Bioscience, Biochemistry and Bioinformatics (IPCBEE '12)*, vol. 3, no. 1, ACSIT Press, Singapoore, 2012.

[33] M. B. Gerstein, A. Kundaje, M. Hariharan et al., "Architecture of the human regulatory network derived from ENCODE data," *Nature*, vol. 489, pp. 91–100, 2012.

[34] D. R. Lovley, T. Ueki T, T. Zhang et al., "Geobacter: the microbe electric's physiology, ecology, and practical applications," *Advances in Microbial Physiology*, vol. 59, pp. 1–100, 2011.

[35] G. Panitchayangkoona, D. V. Voronine, D. Abramavicius et al., "Direct evidence of quantum transport in photosynthetic light-harvesting complexes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, pp. 20908–20912, 2011.

[36] C. Smyth, F. Fassioli, and G. D. Scholes, "Measures and implications of electronic coherence in photosynthetic light-harvesting," *Philosophical Transactions A*, vol. 370, pp. 3728–3749, 2012.

[37] A. Thilagam, "Multipartite entanglement in the Fenna-Matthews-Olson (FMO) pigment-protein complex," *Journal of Chemical Physics*, vol. 136, Article ID 175104, 14 pages, 2012.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 340620, 6 pages
http://dx.doi.org/10.1155/2013/340620

Research Article

State-of-the-Art Fusion-Finder Algorithms Sensitivity and Specificity

Matteo Carrara,[1] Marco Beccuti,[2] Fulvio Lazzarato,[3] Federica Cavallo,[1] Francesca Cordero,[2] Susanna Donatelli,[2] and Raffaele A. Calogero[1]

[1] *Department of Molecular Biotechnology and Health Sciences, University of Torino, Via Nizza 52, 10126 Torino, Italy*
[2] *Department of Computer Science, University of Torino, C.So Svizzera 185, 10149 Torino, Italy*
[3] *Unit of Cancer Epidemiology, Department of Biomedical Sciences and Human Oncology, University of Torino, 10126 Torino, Italy*

Correspondence should be addressed to Raffaele A. Calogero; raffaele.calogero@unito.it

Received 4 October 2012; Revised 11 January 2013; Accepted 15 January 2013

Academic Editor: Tun-Wen Pai

Background. Gene fusions arising from chromosomal translocations have been implicated in cancer. RNA-seq has the potential to discover such rearrangements generating functional proteins (chimera/fusion). Recently, many methods for chimeras detection have been published. However, specificity and sensitivity of those tools were not extensively investigated in a comparative way. *Results*. We tested eight fusion-detection tools (FusionHunter, FusionMap, FusionFinder, MapSplice, deFuse, Bellerophontes, ChimeraScan, and TopHat-fusion) to detect fusion events using synthetic and real datasets encompassing chimeras. The comparison analysis run only on synthetic data could generate misleading results since we found no counterpart on real dataset. Furthermore, most tools report a very high number of false positive chimeras. In particular, the most sensitive tool, ChimeraScan, reports a large number of false positives that we were able to significantly reduce by devising and applying two filters to remove fusions not supported by fusion junction-spanning reads or encompassing large intronic regions. *Conclusions*. The discordant results obtained using synthetic and real datasets suggest that synthetic datasets encompassing fusion events may not fully catch the complexity of RNA-seq experiment. Moreover, fusion detection tools are still limited in sensitivity or specificity; thus, there is space for further improvement in the fusion-finder algorithms.

1. Background

Direct sequencing of messenger RNA transcripts using the RNA-seq protocol [1] is rapidly becoming the standard method for detecting and quantifying genes being expressed in a cell. One of the key features observed when analyzing cancer genomes is chromosomal abnormality. Genome rearrangements could result in aberrant fusion genes, and a number of them have been found to play important roles in carcinogenesis [2]. The discovery of novel gene fusions can lead to a better comprehension of cancer progression and development. The emergence of deep sequencing of transcriptome, known as RNA-seq, has opened many opportunities for the identification of this class of genomic alterations, leading to the discovery of novel chimeric transcripts in many cancers [2]. In this paper, we compare eight fusion-finder softwares to evaluate their relative efficacy in detecting in fusion events.

2. Results

2.1. Fusion Finders. At the best of our knowledge, we have identified Bellerophontes [3], ChimeraScan [4], deFuse [5], FusionFinder [6], FusionHunter [7], FusionMap [8], MapSplice [9], and TopHat-fusion [10] as the most used tools for chimeras detection.

The tools can be organized in various subgroups on the basis of their alignment strategies. In this paper, we propose the following classification: *Whole paired-end, Paired-end + fragmentation*, and *Direct fragmentation*.

In the *Whole paired-end* approach, tools align the full-length paired-end reads on a reference and use discordant

alignments to generate a set of putative fusion events which are finally selected using several additional pieces of information or filtering steps.

Instead, the tools in the *Paired-end + fragmentation* class derive the putative fusion events in two steps. First, as in the *Whole paired-end* approach, the full-length paired-end reads are aligned on a reference, and the discordant alignments are used to generate new pseudoreference including only the identified putative fusion events. Then, reads unaligned in the first step are fragmented and realigned on the pseudoreference to identify junction-spanning reads. Only the putative fusion events associated with junction-spanning reads are selected as input to the filtering step.

Finally, the tools based on *Direct fragmentation* do not directly exploit paired-end information; they fragment every read before the first alignment and find fusion candidates aligning those fragments to a genomic reference.

The putative fusion events are then selected implementing a set of filtering steps.

According to the previous classification the eight tools compared in this paper can be grouped as following:

(1) deFuse and FusionHunter are *Whole Paired-end* based tools;

(2) TopHat-fusion, ChimeraScan, and Bellerophontes are *Paired-end + fragmentation* based tools;

(3) MapSplice, FusionMap, and FusionFinder are *Direct fragmentation* based tools.

Since all the considered tools implement a set of filters to reduce the number of false positive fusion events, a brief description of these filters is reported.

Paired-End Information Filter. It uses the distance between the tags of a pair to validate the alignment on a fusion.

Anchor Length Filter. Anchor length is an important metric for quality evaluation of a read spanning across a fusion junction, and it is defined as the number of nucleotides overlapping each side of the break point. The filter removes all the junction-spanning reads having the anchor length lower than a threshold.

Read-Through Transcripts Filter. It removes the RNA molecules formed by exons of adjacent genes, usually generated by the RNA-polymerase failing the recognition of the gene end.

Junction-Spanning Reads Filter. It discards all fusion events supported by a number of junction-spanning reads lower than a threshold.

PCR-Artifact Filter. It identifies and removes all duplicated reads introduced by the PCR amplification process.

Homology-Based Filter. It removes candidate fusions with a high number of reads on homologous or repetitive regions.

Quality-Based Filters. It is a group of filters that uses different metrics (e.g., entropy, base quality, etc.) for computing the fusion quality. Then, all the candidates with quality lower than a threshold are removed.

Table 1: Filtering steps embedded in the algorithms.

Filters	Fusion finders							
	FF	THF	MS	FM	FH	DF	BF	CS
Pair distance	X					X	X	X
Anchor length		X			X			X
Read-through	X	X		X	X		X	
Junction-spanning				X	X		X	
PCR artifact				X	X		X	
Homology	X	X					X	
Quality			X	X				

FF: FusionFinder; THF: TopHat-fusion; MS: MapSplice; FM: FusionMap; FH: FusionHunter; DF: deFuse; BF: Bellerophontes; CS: ChimeraScan.

In Table 1, we report the implemented filters for each considered tool.

2.2. Fusion Detection Sensitivity. To compare the sensitivity of chimera finder algorithms, we used three datasets.

The first dataset is synthetic (*FM_set*), while the other two are based on real data (*Edgren_set* [11] and *Berger_set* [12]). All the datasets are paired-end ones. The synthetic set is composed of 75 nts reads, while the other two contain 50 nts reads. *FM_set* encompasses 50 fusion events, supported by 9 to 8852 paired-end reads. The *Edgren_set* encompasses a total of 27 experimentally validated fusion genes, detected in BT-474, SK-BR-3, KPL-4, and MCF-7 breast cancer cell lines [11]. *Berger_set* encompasses a total of 12 experimentally validated fusion genes, detected in 501 Mel (Melanoma), K-562 (Leukemia) cell lines, and in 5 samples from primary human melanoma [12].

In this analysis of sensitivity, we considered three parameters: (i) the total number of true positive fusions detected by the different tools (called *all*), (ii) the number of true positive fusions detected by the correct orientation of the two genes (called *right*), and (iii) the number of true positive fusions detected by erroneous orientation of the two genes (called *wrong*).

Using the synthetic *FM_set*, five out of eight analyzed tools show a good sensitivity, since they detect 40 out of 50 fusions (Figure 1(a), blue bars). ChimeraScan was the least sensitive detecting only nine out of 50 fusions (Figure 1). FusionFinder and ChimeraScan were the only tools that did not make any error in the detection of the fusions orientation (Figure 1(a), red bars). It is notable that Bellerophontes was calling all fusion events in both possible orientations, essentially leaving to further down-stream analysis the definition of the correct orientation of fusion events.

The same analysis performed on the *Edgren_set* provided a completely different view of sensitivity of the analyzed tools (Figure 2). From this analysis, ChimeraScan performed better than all the other tools concerning the number of detected chimeras and the correct orientation of the fusion events; 19 out of 27 fusions were all detected in the right orientation. TopHat-fusion was as sensitive as ChimeraScan

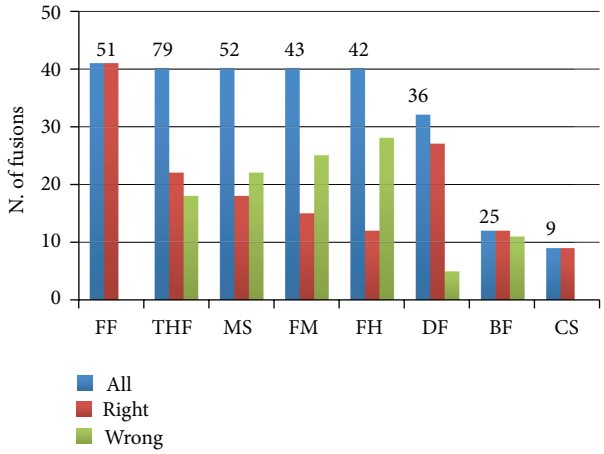


Figure 1: *Fusion events detection performances on positive data set encompassing 50 synthetic fusion events* (*FM_set*). Total number of detected fusions is shown on the top of each bar set. FF: FusionFinder; THF: TopHat-fusion; MS: MapSplice; FM: FusionMap; FH: FusionHunter; DF: defuse; BF: Bellerophontes; CS: ChimeraScan.

detecting 19 chimeras, but only 8 of them were in the correct orientation. Furthermore, the 19 true fusions were part of a set of more than 130000 events, which makes quite difficult the task of purging the false positives. deFuse and FusionFinder came in sensitivity after ChimeraScan and TopHat-fusion, detecting 16 and 13 chimeras, respectively. FusionHunter and FusionMap performance was very poor, with 8 and 4 detected chimeras. MapSplice and Bellerophontes data could not be collected because, after 10 days from the beginning of the analysis, the tools were still filtering fusions events.

We also evaluated the level of overlap between the various tools for the *Edgren_set* chimeras (Figures 2(b) and 2(c)). ChimeraScan encompasses all genes detected by FusionMap and the majority of the fusions detected by the other tools.

Another interesting point is the strong difference between tools in the number of fusions called. At the two extremes are TopHat-fusion, calling more than 130000 chimeras, and FusionHunter, calling only 26 chimeras. We also observed that the best two tools, ChimeraScan and TopHat-fusion, are the ones with the highest number of called fusions. The number of called chimeras is, however, not proportional with the number of detected true positives; for example, both ChimeraScan and TopHat-fusion detect 19 true positives. However, the number of chimeras detected by TopHat-fusion is approximately ten times greater than those detected by ChimeraScan (Figure 2).

We further confirm that ChimeraScan performs better than the other tools also on *Berger_set* (Figure 3). It is notable that the fusion discovery sensitivities for FusionMap, FusionHunter, deFuse, TopHat-fusion, and ChimeraScan, previously observed in the *Edgren_set*, are also kept in the *Berger_set*, and TopHat-fusion is again the best tool in sensitivity after ChimeraScan. However, it is notable that also in *Berger_set*, we have a very high number of called fusions for ChimeraScan and TopHat-fusion, which may make their use in a real experimental setting unpractical.

2.3. False Discovery of Fusions. As shown in the previous paragraph, real datasets are useful to test tools in conditions that resemble their everyday usage. However, real datasets have the limitation that the exact number of true positive fusions is not known; thus, false positive detection cannot be assessed. For this reason, we have used a negative data set (called *negative_set*) encompassing 70 million reads 2×50 nts [13].

FusionHunter and Bellerophontes are the only tools not detecting false chimeras in the negative dataset (Figure 4). The number of false positives increases progressively from FusionMap, deFuse, ChimeraScan, FusionFinder, and MapsSplice to TopHat-fusion, which has the highest number of false positive detected chimeras.

We try to evaluate, for ChimeraScan, if there is a bias in the discovery approach of the tool, which could lead to find the same fusions in different datasets. Intersecting the fusions detected in the *Edgren_set* and in the *Berger_set* and those detected in the *negative_set,* the overlap is marginal. Sixty fusions are found in common between the *negative_set* and the *Edgren_set*, 197 between the *negative_set* and the *Berger_set,* and only 38 fusions are in common between the previous two comparisons. This observation suggests that false positives are mainly dataset specific and not significantly biased by the tool characteristics.

2.4. Optimizing Removal of False Positive Fusions. Being ChimeraScan the most efficient tool in detection of fusion events in the right orientation, we evaluated various filtering approaches to reduce the false positive fusions contaminating the real fusion events. Specifically, we used the characteristics of the chimeras detected in the *negative_set* to define false positive filters. We observed that many chimeras found in the negative set by ChimeraScan were supported by reads encompassing the two exons of the genes involved in the fusion and were missing reads spanning over the fusion junction. Since the presence of junction-spanning reads is an important positive parameter for the definition of a fusion event [2], we decided to filter-out all the chimeras detected in the *Edgren_set* but not supported by reads spanning over the fusion junction. The filter is very efficient since we retain only 681 fusions out of the initial 13346 detected by ChimeraScan. Concerning the true positives, 17 out of the 19 detected fusions are also retained. RPS6KB1:SNF8 and CPNE1:PI3 are instead lost.

It is interesting to note that RPS6KB1:SNF8 can be detected by deFuse, FusionHunter, and TopHat-fusion, while CPNE1:PI3 could be found by FusionFinder and TopHat-fusion. All the previously mentioned methods manage to detect spanning reads for RPS6KB1:SNF8 and CPNE1:PI3, suggesting that ChimeraScan algorithm fails to detect those spanning reads. We are currently trying to find out the reason why ChimeraScan failed in detecting the prviously mentioned fusion junction spanning reads. Furthermore, tools already implementing a filter based on the number of junction-spanning reads consistently show a lower number of reported fusions.

We have also observed the presence of a high number of fusions encompassing intronic region in the fusions detected

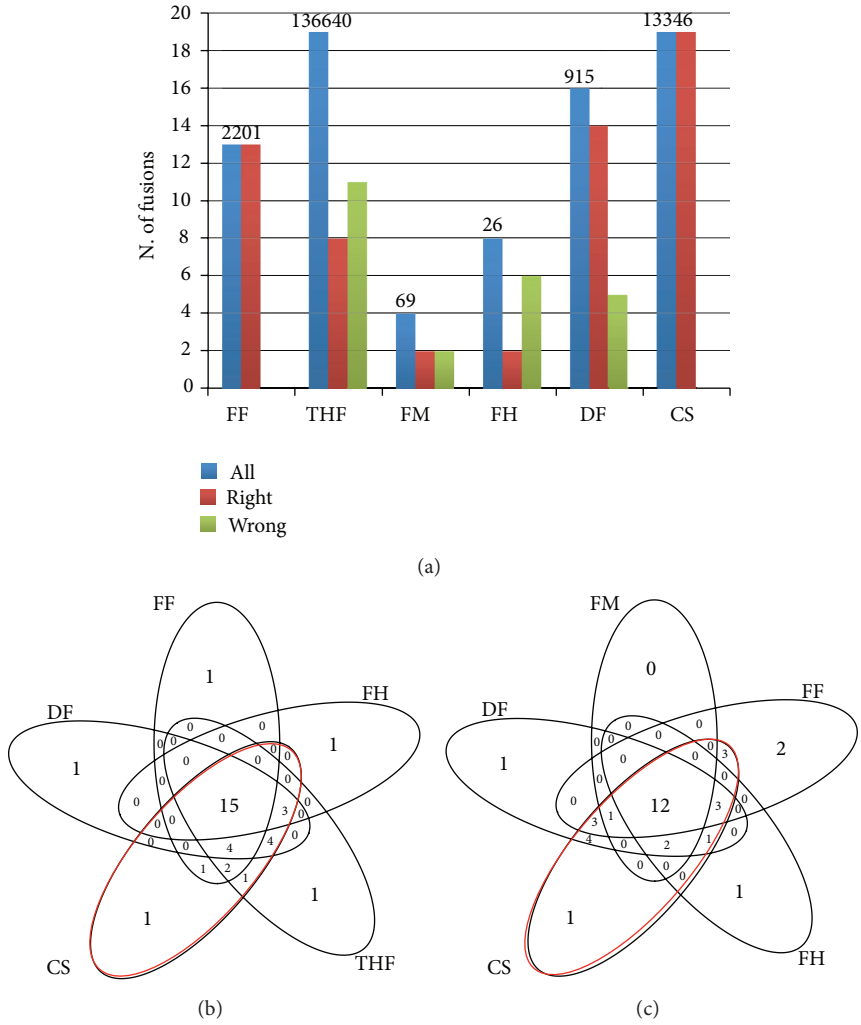


Figure 2: *Analysis of sensitivity of fusion finders in a real data set encompassing 27 validated fusions* (*Edgren_set*). (a) Total number of detected fusions is shown on the top of each bar set. (b) and (c) Venn diagrams showing the overlaps between fusions founded by different tools. The ellipse of the ChimeraScan is highlighted in red. FF: FusionFinder; THF: TopHat-fusion; MS: MapSplice; FM: FusionMap; FH:FusionHunter; DF: defuse; BF: Bellerophontes; CS: ChimeraScan.

in *negative_set*. These fusions generate very large transcripts, which do not produce in frame transcripts. Removing them from the 681 fusions detected in the *Edgren_set,* we retain 249 chimeras, without loss of true positives. Again, some of tools include alignment approaches with an effect similar to this filter by aligning reads to the transcriptome.

Although 249 chimeras represent a significant reduction of the initial number of detected chimeras, they are still too many to be all experimentally validated. Sorting the 249 chimeras in descending order, on the basis of the number of fusion junction-spanning reads, we show that with the top 17 chimeras, 10 were part of the 17 true positives. The rationale of this ranking procedure is that biological effect also depends on the amount of the expressed mRNA; thus, highly expressed fusions, that is, fusions with a high number of junction-spanning reads, might have a more important role in cancer physiology.

3. Discussion

The main goal of this paper is to understand strength and limits of the main fusion detection software currently available. To reach our aim, we have evaluated sensitivity and false fusion discovery for eight state-of-the-art fusion finders: Bellerophontes, FusionHunter, FusionMap, FusionFinder, MapSplice, deFuse, ChimeraScan, and TopHat-fusion. We run this comparison using both synthetic and real datasets.

Concerning sensitivity, we observed that a comparison analysis run only on synthetic data could generate misleading results. Sensitivity analysis run on the synthetic data only results in ChimeraScan being the least sensitive tool, while it is actually the most sensitive tool on real datasets. We think that discrepancies between results obtained on synthetic and real data are due to the actual lack of knowledge of the real complexity of RNA-seq data that does not allow

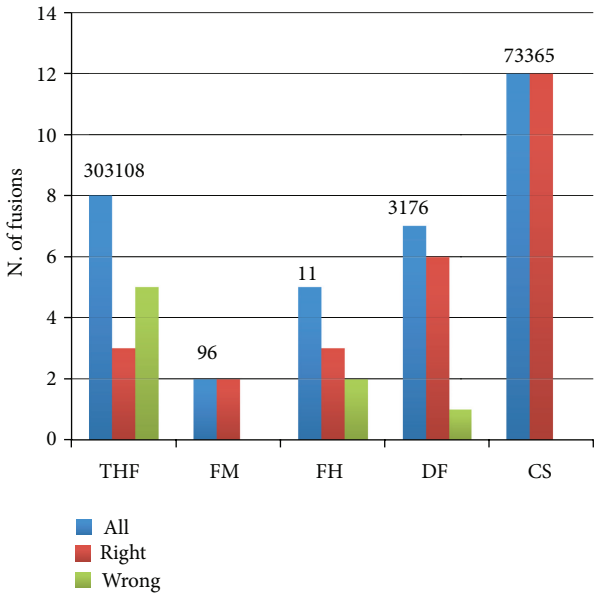


Figure 3: *Analysis of sensitivity of fusion finders on a real data set encompassing 12 validated fusions* (*Berger_set*). Total number of detected fusions is shown on the top of each set of bars. FM: FusionMap; FH: FusionHunter; DF: defuse; CS: ChimeraScan; THF: TopHat-fusion.

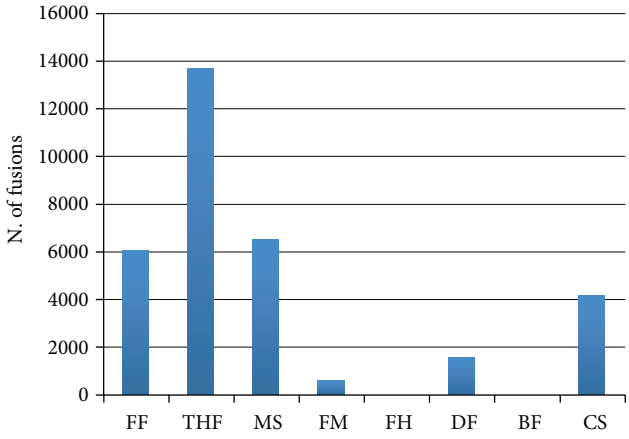


Figure 4: *False positive fusion detected using a synthetic dataset without chimeras.* FF: FusionFinder; THF: TopHat-fusion; MS: MapSplice; FM: FusionMap; FH: FusionHunter; DF: defuse; BF: Bellerophontes; CS: ChimeraScan.

the construction of fully significant synthetic datasets. The analysis of real datasets allows us to identify ChimeraScan as the most sensitive tool for chimeras detection although ChimeraScan output is affected by a very high number of called fusions, a number too big to make a functional experimental validation feasible. A synthetic dataset, free of fusion events by construction (*negative_set*), represents an interesting instrument to understand the basic characteristics of false fusions detected by ChimeraScan and to define specific filters to remove them. It was observed that the main characteristics of false positive fusions in the *negative_set* are both the lack of fusion junction-spanning reads and the inclusion of intronic regions in the fusion. The application of two filters, based on the previous false positive characteristics, proves to be very efficient in reducing the 13346 initially detected fusions (*Edgren_set*) to 249, with the limited loss of two true positive chimeras. It is also notable that ranking chimeras on the basis of the number of fusion junction-supporting reads also helps to further narrow the set of chimera to be experimentally validated.

4. Conclusions

This paper highlights that fusion detection tools are still not fully adequate to provide a direct solution for the discovery of chimeras in a dataset. Many algorithms have been proposed, and each of them has specific biases at the level of sensitivity or specificity. Tools having low sensitivity are also characterized by a limited number of false positives. Moreover, results obtained by the low sensitivity tools show very limited overlap in the results. On the other hand, tools as ChimeraScan and TopHat-fusion show a good sensitivity but also the presence of a high number of false positives. Filters devoted to the removal of false positives can significantly improve the ratio between true positives and false positives, but there is clearly space for algorithm improvements.

5. Methods

5.1. Fusion Detection Softwares and Data Analysis. FusionHunter, FusionMap, FusionFinder, MapSplice, deFuse, ChimeraScan, Bellerophontes, and TopHat-fusion were downloaded from the repositories indicated in their publications and installed following requirements indicated in their manuals. Software was run using default configuration. All analyses were performed on a 48-core AMD server with 512 Gb RAM and 9 Tb HD, running linux SUSE Enterprise 11. Statistics and data parsing were executed using R scripting, taking advantage of Bioconductor [14] packages, that is, Biostrings, org.Hs.eg.db, GenomicRanges, and oneChannelGUI [15].

5.2. Positive Dataset. FusionMap developers provide a synthetic dataset of simulated paired-end RNA-seq reads (~60,000 pairs of reads, 75 nt, fragment size = 158 bp). Fifty fusions are represented with a range of supporting pairs going from 9 to 8852. Real datasets encompassing experimentally validated chimeras were retrieved from NCBI Sequence Read Archive (SRA:SRP003186) as described in [11] and from NCBI Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/), SuperSeries Accession no. GSE17593 as described in [12].

5.3. Negative Dataset. The negative dataset was generated using BEERS [16], and its construction is described in [13].

Authors' Contribution

F. Lazzarato installed and set up fusions detection software and databases. M. Carrara and M. Beccuti performed the

comparison among fusion finders. R. A. Calogero collected data and generated negative dataset. F. Cavallo and S. Donatelli revised the paper and provided suggestions. R. A. Calogero and F. Cordero supervised the overall work. These authors contributed equally to this work.

Acknowledgments

This study was funded by grants from the Italian Association for Cancer Research; the Epigenomics Flagship Project EPIGEN, MIUR-CNR; the Italian Ministero dell'Università e della Ricerca; the University of Torino and Regione Piemonte; FP7-Health-2012-Innovation-1 NGS-PTL Grant no. 306242. The work of M. Beccuti has been supported by project Grant no. 10-15-1432/HICI from the King Abdulaziz University of Saudi Arabia.

References

[1] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by RNA-Seq," *Nature Methods*, vol. 5, no. 7, pp. 621–628, 2008.

[2] C. A. Maher, C. Kumar-Sinha, X. Cao et al., "Transcriptome sequencing to detect gene fusions in cancer," *Nature*, vol. 458, no. 7234, pp. 97–101, 2009.

[3] F. Abate, A. Acquaviva, G. Paciello et al. et al., "Bellerophontes: an RNA-Seq data analysis framework for chimeric transcripts discovery based on accurate fusion model," *Bioinformatics*, vol. 28, no. 16, pp. 2114–2121, 2012.

[4] M. K. Iyer, A. M. Chinnaiyan, and C. A. Maher, "ChimeraScan: a tool for identifying chimeric transcription in sequencing data," *Bioinformatics*, vol. 27, no. 20, pp. 2903–2904, 2011.

[5] A. McPherson, F. Hormozdiari, A. Zayed et al. et al., "deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data," *PLoS Computational Biology*, vol. 7, no. 5, Article ID e1001138, 2011.

[6] R. W. Francis, K. Thompson-Wicking, K. W. Carter, D. Anderson, U. R. Kees, and A. H. Beesley, "FusionFinder: a software tool to identify expressed gene fusion candidates from RNA-Seq data," *PloS One*, vol. 7, no. 6, Article ID e39987, 2012.

[7] Y. Li, J. Chien, D. I. Smith, and J. Ma, "FusionHunter: identifying fusion transcripts in cancer using paired-end RNA-seq," *Bioinformatics*, vol. 27, no. 12, Article ID btr265, pp. 1708–1710, 2011.

[8] H. Ge, K. Liu, T. Juan, F. Fang, M. Newman, and W. Hoeck, "FusionMap: Detecting fusion genes from next-generation sequencing data at base-pair resolution," *Bioinformatics*, vol. 27, no. 14, Article ID btr310, pp. 1922–1928, 2011.

[9] K. Wang, D. Singh, Z. Zeng et al., "MapSplice: accurate mapping of RNA-seq reads for splice junction discovery," *Nucleic Acids Research*, vol. 38, no. 18, p. e178, 2010.

[10] D. Kim and S. L. Salzberg, "TopHat-Fusion: an algorithm for discovery of novel fusion transcripts," *Genome Biology*, vol. 12, no. 8, article R72, 2011.

[11] H. Edgren, A. Murumagi, S. Kangaspeska et al., "Identification of fusion genes in breast cancer by paired-end RNA-sequencing," *Genome Biology*, vol. 12, no. 1, article no. R6, 2011.

[12] M. F. Berger, J. Z. Levin, K. Vijayendran et al., "Integrative analysis of the melanoma transcriptome," *Genome Research*, vol. 20, no. 4, pp. 413–427, 2010.

[13] M. Carrara, M. Beccuti, F. Cavallo et al., "State of art fusion-finder algorithms are suitable to detect Transcription-Induced Chimeras in normal tissues?" *BMC Bioinformatics*. In press.

[14] R. C. Gentleman, V. J. Carey, D. M. Bates et al., "Bioconductor: open software development for computational biology and bioinformatics," *Genome Biology*, vol. 5, no. 10, article R80, 2004.

[15] R. Sanges, F. Cordero, and R. A. Calogero, "oneChannelGUI: a graphical interface to Bioconductor tools, designed for life scientists who are not familiar with R language," *Bioinformatics*, vol. 23, no. 24, pp. 3406–3408, 2007.

[16] G. R. Grant, M. H. Farkas, A. D. Pizarro et al., "Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM)," *Bioinformatics*, vol. 27, no. 18, pp. 2518–2528, 2011.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 802945, 9 pages
http://dx.doi.org/10.1155/2013/802945

Research Article

On the Structural Context and Identification of Enzyme Catalytic Residues

Yu-Tung Chien and Shao-Wei Huang

Department of Medical Informatics, Tzu Chi University, 701 Zhongyang Road, Section 3, Hualien 97004, Taiwan

Correspondence should be addressed to Shao-Wei Huang; swhwang.orz@gmail.com

Received 29 November 2012; Accepted 28 December 2012

Academic Editor: Tun-Wen Pai

Enzymes play important roles in most of the biological processes. Although only a small fraction of residues are directly involved in catalytic reactions, these catalytic residues are the most crucial parts in enzymes. The study of the fundamental and unique features of catalytic residues benefits the understanding of enzyme functions and catalytic mechanisms. In this work, we analyze the structural context of catalytic residues based on theoretical and experimental structure flexibility. The results show that catalytic residues have distinct structural features and context. Their neighboring residues, whether sequence or structure neighbors within specific range, are usually structurally more rigid than those of noncatalytic residues. The structural context feature is combined with support vector machine to identify catalytic residues from enzyme structure. The prediction results are better or comparable to those of recent structure-based prediction methods.

1. Introduction

Understanding the molecular mechanisms of enzyme catalysis is important in studies of various complicated biological processes. The number of protein structures deposited to the Protein Data Bank [1] has increased rapidly in the past decade. However, the function and catalytic residues of a large fraction of enzymes are not well studied and understood. Experimental methods which are used to identify enzyme catalytic residues, like site-directed mutagenesis, are time consuming and expensive. Computational methods designed to identify catalytic residues are needed to efficiently handle the huge number of proteins whose catalytic sites are not determined.

Many methods have been developed to predict protein catalytic sites based on information extracted from protein seqence and structure. One of the most direct strategies is based on finding homologous enzymes whose function and catalytic residues are already known [2–6]. Catalytic residues of a novel protein are identified by using sequence or structure similarity search with enzymes whose catalytic residues were well annotated. However, there are still limitations for such methods based on homology search. First, homologous enzymes whose function and catalytic sites are already known are needed. Second, proteins of similar tertiary structures do not always have completely identical function [7]. There are also examples showing that proteins of different tertiary structures have the same function [8].

To directly identify catalytic sites from single protein structure without needing homology information, it is important to study the fundamental differences between catalytic residues and noncatalytic residues. Sacquin-Mora et al. [9] used the computation of a force constant, that is, the ease of moving a given residue with respect to the other residues in the protein, to identify catalytic residues and found that the catalytic residues usually have higher force constant. Ben-Shimon and Eisenstein [10] found that the catalytic residues are often located near the small fractions of the exposed residues closest to the center of the protein. Amitai et al. [11] converted protein to a network in which the residues are vertices and their interactions are edges and showed that the central hubs in the network are usually functional important residues or residues having direct contact with them. Wie et al. [12] developed a method, Theoretical Microscopic Titration Curves (THEMATICS), which computes residue electrostatic properties from protein

structure, to identify catalytic residues. The THEMATICS method was then combined with geometry features derived from protein structure to predict catalytic residues from enzyme structure using a monotonicity-constrained maximum likelihood approach, called Partial Order Optimum Likelihood (POOL) [13]. A more recent method, EXIA [14], successfully identifies catalytic residues based on residue side chain orientation of single enzyme structure without needing structure or sequence homology information.

In this study, we first analyzed the structural context of catalytic and noncatalytic residues based on their sequence and structure neighbors. We show that catalytic residues are usually located in structurally more rigid environment than noncatalytic residues. The sequence or structural neighboring residues within specific range of catalytic residues have distinct structural features. We further combined the structural context features and support vector machine to identify catalytic residues from protein structure.

2. Methods

2.1. Calculation of Structural Context. The weighted-contact number model (WCN) [15, 16] is used to calculate structural flexibility of residue environment. WCN is highly correlated to experimental B-factor and order parameter of protein structure solved by nuclear magnetic resonance. The WCN of the ith residue is based on the distances between the ith residue and all the other residues in the enzyme, as in

$$D_i = \sum_{j \neq i}^{N} \frac{1}{r_{ij}^2}, \tag{1}$$

where N is total number of residues in the enzyme, and r_{ij} is the distance between ith and jth residues. The coordinate of Cα atom is used to represent the position of the residue.

There are two types of structural context: sequence neighbor flexibility (SEQ) and structure neighbor flexibility (STR). The SEQ of the ith residue is defined as the average structural flexibility of the ith residue and its flanking residues on sequence as in

$$\mathrm{SEQ}_i = \frac{\sum_{x=i-n}^{x \leq i+n} D_x^{-1}}{(2n+1)}, \tag{2}$$

where residues $i-n$ to $i+n$ are the nearest n neighbors of the ith residue on sequence. WCN is inverted for an easy comparison with B-factor. If x is out of the range of the sequence, it is simply ignored in the calculation. The STR of the ith residue is defined as the average structural flexibility of the ith residue and residues whose distance to the ith residue are smaller than a cut-off value as in

$$\mathrm{STR}_i = \frac{\sum_{x \in M} D_x^{-1}}{m}, \tag{3}$$

where M is a subset of residues whose distance to the ith residue is smaller than the cut-off distance and m is the number of residues in the subset. The concept of SEQ and STR is extended from our previous work [17], which only considered the nearest two sequence neighbors.

2.2. Normalization of Structural Context and B-Factor Profiles. The SEQ, STR, and B-factor are normalized to their corresponding z-scores:

$$z_x = \frac{x - \overline{x}}{\sigma_x}, \tag{4}$$

where $\overline{x}$ and σ_x are the mean and standard deviations of x of a given protein. In this work, x is SEQ, STR, or B-factor from X-ray crystallographic structures. For a given protein, the mean and standard deviations are calculated based on the scores of the protein. The scores of the protein are then normalized according to its mean and standard deviations. The normalized SEQ, STR, and B-factor are referred to as Z_{SEQ}, Z_{STR}, and Z_{B}, respectively. For convenience, the normalized SEQ, STR, and normalized B-factor profiles are simply called SEQ profile, STR profile and B-factor profile.

2.3. The Support Vector Machine. SVM finds the separating hyperplane with the largest distance between two classes. However, the data being classified may not always be linearly separable in the space. It was proposed that the original space be mapped into a higher-dimensional space, making the separation easier in that space. The support vector machine method (SVM) has been widely applied to many bioinformatics studies: protein-fold assignment [18, 19], subcellular localization prediction [20, 21], secondary-structure prediction [22–24], and other biological pattern-classification problems [25–28]. SVMs perform well in these classification problems when compared to other machine-learning methods because of thier convenient classifier's capacity control and avoidance of overfitting. In this work, the software package LIBSVM [29] version 3.11 was used.

Here we used SVM to predict catalytic residues using the structural context features, SEQ and STR, as input features. The feature vector for a residue is one of these features or their combinations: Z_{SEQ}, Z_{STR}, Z_{B}, or binary coding of amino acid type. The common problem encountered in enzyme catalytic site prediction using SVM is the extremely unbalanced ratio of catalytic residues and noncatalytic residues. A well-used strategy is to randomly select subsets which have a balanced ratio between catalytic and noncatalytic residues when training [30]. Here, a 5-fold cross-validation procedure was used for performance measurement. For each fold, the training data was a randomly selected balanced subset of residues by subsampling noncatalytic residues. The *cost* and *gamma* are parameters used in model training and kernel function of LIBSVM and need to be to tuned for optimal prediction results. These parameters are tuned independently using 5-fold cross-validation for each training dataset. In addition to cost and gamma parameters, other settings used in the SVM include the type of SVM: C-SVC; the type of kernel function: radial basis function. Other parameters not mentioned here are set as their default value in the LIBSVM software.

2.4. Sequence Conservation Score. For comparison with the POOL method, sequence conservation which includes evolutionary information is used as training feature in some

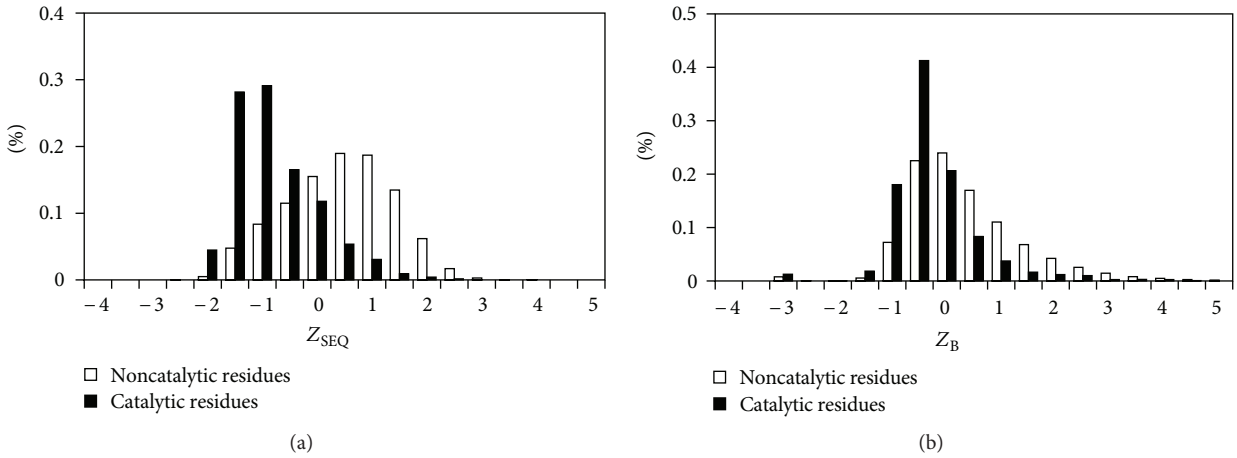


Figure 1: Distributions of (a) Z_{SEQ} and (b) Z_B profiles of catalytic and noncatalytic residues for the E760 dataset.

predictions reported here. Sequence conservation is from position-specific substitution matrix (PSSM) generated by PSI-Blast [31] for each protein. PSI-Blast is set to search against the nonredundant (nr) database for three iterations with default E-value threshold of 5×10^{-3}. The sequence conservation score is directly taken from the "information per position" column in the PSSM profile.

2.5. Dataset. The dataset used in this work is collected from Catalytic Site Atlas (CSA) [32] version 2.2.10 using BlastClust [31]. The dataset contains 760 proteins with pairwise sequence identity ≤30%, including a total of 592,382 residues in which 2,355 residues are catalytic sites. All heteroatoms, ligands, and nonprotein molecules are removed. The dataset is referred to as E760 dataset.

2.6. Evaluation of Prediction Performance. Sensitivity, specificity, and Matthew's correlation coefficient (MCC) were used for performance measure as follows:

$$\text{Sensitivity} = \frac{\text{TP}}{(\text{TP} + \text{FN})},$$

$$\text{Specificity} = \frac{\text{TN}}{(\text{TN} + \text{FP})},$$

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP}+\text{FP}) \times (\text{TP}+\text{FN}) \times (\text{TN}+\text{FN}) \times (\text{TN}+\text{FP})}}, \quad (5)$$

where TP, FP, TN, and FN are the number of true positive, false positive, true negative, and false negative, respectively. A catalytic residue is either TP when correctly predicted to be catalytic residue or FP when incorrectly predicted to be noncatalytic residue. A noncatalytic residue is either TN when correctly predicted to be noncatalytic residue or FN when incorrectly predicted to be catalytic residue. We used MCC to evaluate the performances because MCC takes into account true and false positives and negatives and is a balanced measure especially when the numbers of positives (catalytic residues) and negatives (noncatalytic residues) are extremely unbalanced. Note that the MCC, sensitivity, and specificity reported here are based on balanced data, that is, the numbers of catalytic and noncatalytic residues are equal. They were only used to compare the results between different features in this paper but not used to compare with other methods. The Receiver Operating Characteristic (ROC) curve was calculated based on unbalanced data and was used to compare prediction results with other methods. The ROC curve was plotted by averaging per-protein ROC curve as used in [13].

3. Results and Discussions

First, we discuss the distributions of Z_{SEQ}, Z_{STR}, and Z_B of catalytic residues and noncatalytic residues for the E760 dataset. Then we show the prediction results based on Z_{SEQ} profile, Z_B profile, and amino acid type. Finally, we compared the prediction results based on Z_{SEQ} with those of the methods using other structure-based features.

3.1. Distributions of SEQ for Catalytic and Noncatalytic Residues. In this section, we compare the distributions of SEQ (sequence neighbor flexibility) for catalytic residues and noncatalytic residues. Figure 1(a) displays the distributions of Z_{SEQ} when $n = 1$ (n: the number of flanking neighboring residues on sequence to calculate the average structural flexibility) for catalytic residues and noncatalytic residues for the E760 dataset. For comparison, the distributions of Z_B are also shown in Figure 1(b). The distributions of Z_{SEQ} for catalytic and noncatalytic residues show that catalytic residues are much less flexible and located in a more rigid context than noncatalytic residues. The phenomenon is much more significant using Z_{SEQ} than using Z_B as shown in Figure 1. There are 90% of catalytic residues having $Z_{SEQ} \leq 0$ and 40% of noncatalytic residues having $Z_{SEQ} \leq 0$. Only 81% of catalytic residues have $Z_B \leq 0$ and 54% of noncatalytic residues have $Z_B \leq 0$. SEQ and crystallographic B-factor are both based on the number and distances of neighbors

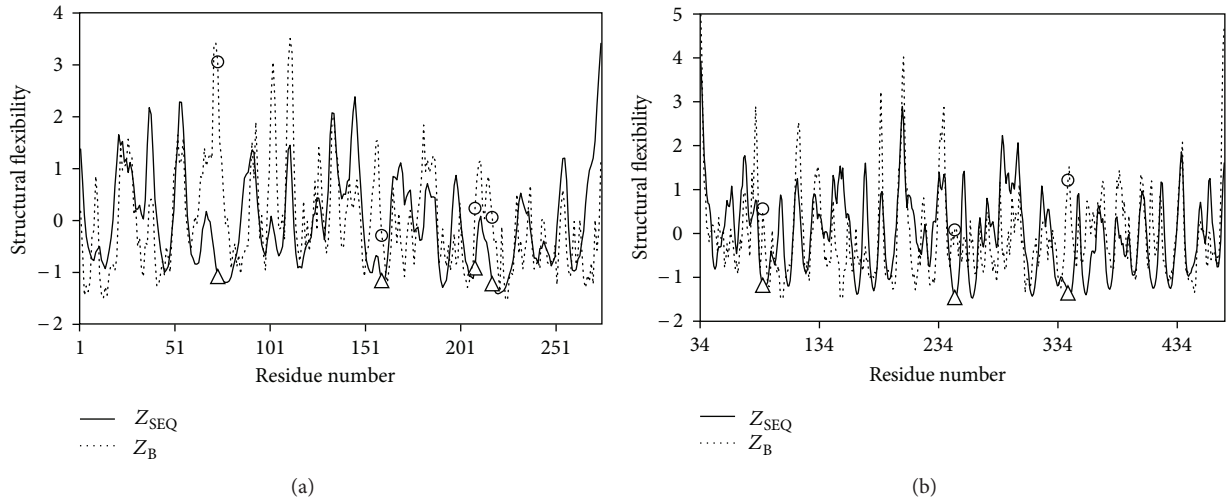


FIGURE 2: Comparison of SEQ and B-factor profiles for two example proteins. Comparison of Z_{SEQ} and Z_B profiles for (a) diaminopimelate epimerase (PDB: 1BWZ) and (b) levansucrase (PDB: 1OYG). The catalytic residues are labeled as triangles on Z_{SEQ} profile and as circles on Z_B profile. The catalytic residues of 1BWZ are Cys73, His159, Glu208, and Cys217. The catalytic residues of 1OYG are Asp86, Asp247, and Glu342.

around a given residue and are related to structural flexibility. The SEQ profile is a better and more reliable characteristic to identify catalytic residues than the B-factor profile. B-factor is easily affected by experimental conditions, crystal packing, existence of ligands, or temperature. Two structures of the same enzyme under different experimental conditions may have very different B-factor profiles but have almost identical crystal structures.

Figure 2(a) shows the Z_{SEQ} of $n = 1$ (solid lines) and Z_B (dashed lines) profiles of two enzymes, diaminopimelate epimerase (PDB id: 1BWZ) and levansucrase (PDB id: 1OYG). The catalytic residues and noncatalytic residues are labeled as triangle and circle on Z_{SEQ} profiles and on Z_B profiles, respectively. It is obvious that the catalytic residues are located in the most structurally stable regions for the Z_{SEQ} profiles in both examples. The four catalytic residues (Cys73, His159, Glu208, and Cys217) of diaminopimelate epimerase are located near the centroid of the enzyme, forming a rigid catalytic spot. Cys73 and Cys217 are close to each other and connected by a disulfide bond. However, they have unusually high Z_B (3.06 and 0.06, resp.) but reasonable low Z_{SEQ} values (−1.08 and −1.22). His159 is partially buried by surrounding neighbors and has a quite low solvent accessible surface (SAS) of 3 Å^2, calculated by the DSSP program [33]. It has a relatively low Z_B (−0.29) and an extremely low Z_{SEQ} (−1.16) comparing to other residues in the enzyme. Glu208 is relatively more exposed to solvent (SAS = 19 Å^2) than His159. It has a high Z_B (0.23) but a very low Z_{SEQ} (−0.91). In the protein, the four catalytic residues are structurally rigid, having very low Z_{SEQ} and SAS values. However, their Z_B are high, especially for Cys73 that forms a disulfide bond with another catalytic residue, Cys217.

The catalytic site of the second example, levansucrase, is constituted of three catalytic residues, Asp86, Asp247, and Glu342, which are inside a cleft near the geometrical center of the protein. They are moderately accessible to solvent (with SAS: 20 Å^2, 8 Å^2, and 24 Å^2, resp.) but are surrounded and thus strongly stabilized by a large number of residues because of their location. The Z_{SEQ} for these three catalytic residues, Asp86, Asp247, and Glu342, are −1.18, −1.46, and −1.36, respectively. Their Z_B values are surprisingly not low (0.56, 0.08, and 1.21, resp.). SEQ is a better estimation of structural flexibility than B-factor whether the location of residue is exposed to solvent or buried inside the protein.

3.2. Distributions of SEQ and STR Based on Different Parameter Settings. In the previous section, we discussed the SEQ when parameter $n = 1$, that is, average WCN of target residue and its nearest two neighboring residues on sequence. Here, we extend the analysis to SEQ calculated based on different window sizes. Figure 3 shows the distributions of SEQ for catalytic and noncatalytic residues with incremental n from 1 to 20. When n is set to 1, the SEQ distributions of catalytic and noncatalytic residues are obviously different. As window size increases, the differences between distributions decrease gradually. The difference between SEQ of catalytic and noncatalytic residues is less obvious when n is larger than 10. Table 2 lists the prediction results when using SEQ with different n. The MCC obviously drops when n is equal or larger than 10. The results indicate that the sequence neighbors of catalytic residues are also structurally more rigid than noncatalytic residues in the range of $n < 10$.

For STR, the cut-off distance for structurally neighboring residues is set from 3 Å to 25 Å as shown in Figure 4. Catalytic residues have structurally rigid neighbors for neighboring residues within 15 Å cut-off distance. When the cut-off distance is larger than 19 Å, catalytic and noncatalytic residues have similar STR distributions. The results suggest that catalytic residues are usually located in structurally stable environments and the surrounding neighboring residues within 15 Å are also relatively structurally rigid.

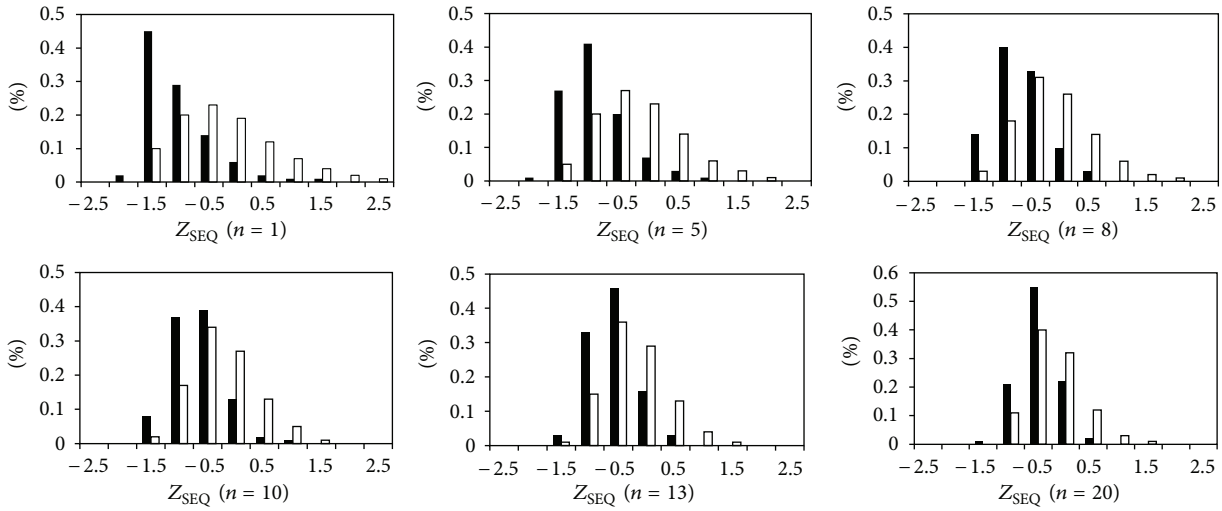


Figure 3: Distributions of Z_{SEQ} for catalytic and noncatalytic residues with different average window sizes.

Table 1: Prediction performances using SVM with different features.

Feature set	Performance		
	Sensitivity	Specificity	MCC
AA[a]	0.70	0.70	0.40
SEQ[b]	0.76	0.70	0.47
STR[c]	0.79	0.69	0.47
SEQ + STR[d]	0.78	0.66	0.45
B-factor	0.63	0.62	0.25
AA + SEQ	0.74	0.76	0.51

[a]Amino acid type.
[b]SEQ with $n = 1$.
[c]STR with cut-off distance = 3 Å.
[d]SEQ with $n = 1$ combined with STR with cut-off distance (3 Å).

3.3. Prediction of Catalytic Residues Based on SEQ and STR. In this section, we discuss the prediction results using SVM with several different features, including amino acid type, SEQ ($n = 1$), STR (cutoff = 3 Å), B-factor profile, and their combinations. The prediction sensitivity, specificity, and MCC using different feature sets are listed in Table 1, including amino acid type (AA), SEQ and STR profiles, B-factor profile (B), combination of amino acid type, and SEQ (AA + SEQ).

The prediction results show that SEQ and STR are much better features than B-factor (MCC = 0.47 for SEQ and STR, 0.25 for B-factor) for identification of catalytic residues. The prediction performances of STR and SEQ are quite similar (sensitivity = 0.76 and 0.79, specificity = 0.70 and 0.69 for SEQ and STR, resp.). We selected SEQ for further-detailed analysis and comparison. Due to the fact that about 95% catalytic residues are polar or charged amino acids, prediction purely based on amino acid type have a MCC of 0.40, which is much higher than that of B-factor. However, the results also show that there are many false positives (specificity = 0.70), which means that catalytic residues have other unique features. SEQ provides information of structural flexibility of residues and their neighbors, which is complementary to amino acid type information. The prediction results that include SEQ and amino acid type show that catalytic residues can be more accurately identified using both features. The MCC is 0.51 and the sensitivity and specificity are 0.74 and 0.76, respectively. The prediction results based on combining SEQ ($n = 1$) and STR (cutoff = 3 Å) are also listed in Table 1. The prediction performance is similar to those based on SEQ or STR alone. The reason may be that the combination of SEQ and STR does not provide more information than SEQ or STR alone, thus not further improving the prediction results.

3.4. Comparison with Structure-Based Prediction Method. To test the performance of SEQ, we compared our prediction results with those of Partial Order Optimum Likelihood (POOL) [6], which combines residue electrostatic properties and structure geometry information to predict catalytic residues. POOL is one of the most successful structure-based prediction methods and it is able to work without needing sequence homology information. First, we directly compared the prediction results of SEQ and those of POOL on a dataset of 160 enzymes [6]. Figure 5 shows the ROC curves of SEQ and POOL based on different features including POOL(T): residue electrostatic properties, POOL(G): structure geometry feature, and POOL(C): sequence conservation. SEQ apparently outperforms POOL(C) and POOL(G) and performs better than POOL(T) and POOL(G+C) (POOL(G) combined with POOL(C)) when false-positive rate is smaller than 0.1. Under higher false-positive rates, POOL(T) and POOL(G + C) have better performance than that of SEQ. It is somewhat interesting that SEQ, which only uses structural rigidity, has comparable results as those of POOL, which uses residue biochemical features, evolutionary sequence conservation, and cleft shape. The results reported here are based

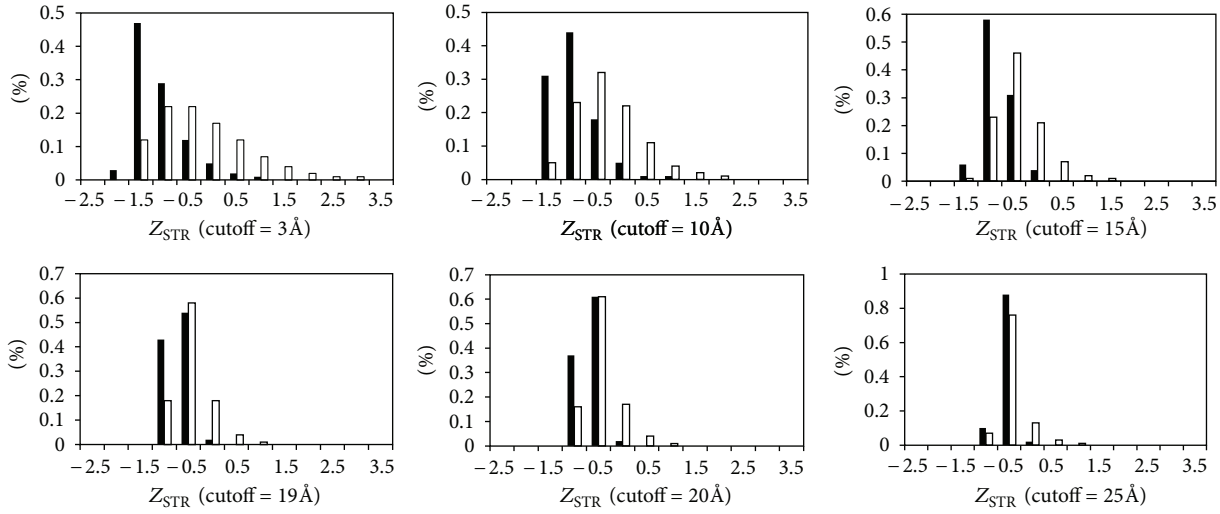


Figure 4: Distributions of Z_{STR} for catalytic and noncatalytic residues with different cut-off distances.

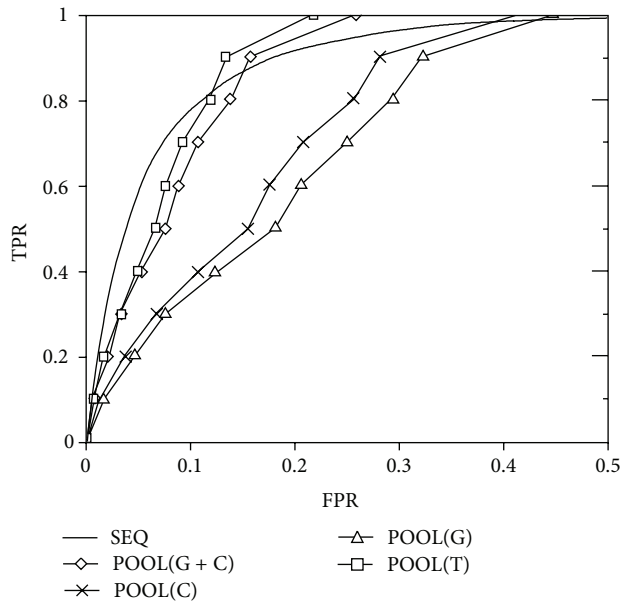


Figure 5: Comparison of ROC curves of SEQ and POOL using different features. The ROC curves are prediction results on a dataset of 160 enzymes. POOL features are denoted as POOL(T): residue electrostatic; POOL(G): structure geometry; POOL(C): sequence conservation; POOL(G + C): structure geometry combined with sequence conservation. The figure was remade from [6].

on fivefold cross-validation on the dataset of 160 enzymes. The ratio between catalytic and noncatalytic residues is not changed (unbalanced) for each test fold.

We also compared the results of SEQ and those of POOL(T + G) (residue electrostatic combined with structure geometry) on a dataset of 79 enzymes [24]. Figure 6 shows the ROC curves of SEQ (dotted line) and those of POOL.

Table 2: Prediction results when using SEQ with different n parameters.

n	Performance		
	Sensitivity	Specificity	MCC
1	0.76	0.70	0.47
5	0.75	0.71	0.47
8	0.79	0.65	0.44
10	0.80	0.57	0.36
13	0.78	0.55	0.36
20	0.77	0.41	0.36

In the results of Figure 5, SEQ performs much better than POOL(G) and is comparable to POOL(T). When POOL(T) and POOL(G) are combined together (POOL(T + G)), it performs better than SEQ (Figure 6). POOL has the best performance when sequence conservation (POOL(C)) is further added (POOL(T + G + C)). To compare with their results, we combined SEQ and sequence conservation by PSI-BLAST. The results show that SEQ performs even better than POOL(T + G + C) when sequence conservation is added (thick solid line in Figure 6). It suggests that although SEQ can find out rigid regions in enzyme structures, amino acid information is still important for the identification of catalytic residues due to the fact that a large fraction of catalytic residues are polar amino acids. Using SEQ without any amino acid information to predict catalytic residue may result in some false positives, for example, rigid but nonpolar residues.

To further compare the results of SEQ and POOL, we have submitted two enzyme structures, diaminopimelate epimerase (PDB code: 1BWZ) and levansucrase (PDB code: 1OYG), to the POOL webserver. For diaminopimelate epimerase, the four catalytic residues, Cys73, His159, Glu208, and Cys217, are ranked 18, 3, 6, and 59, respectively, by POOL (residues higher ranked in POOL are more probable to be

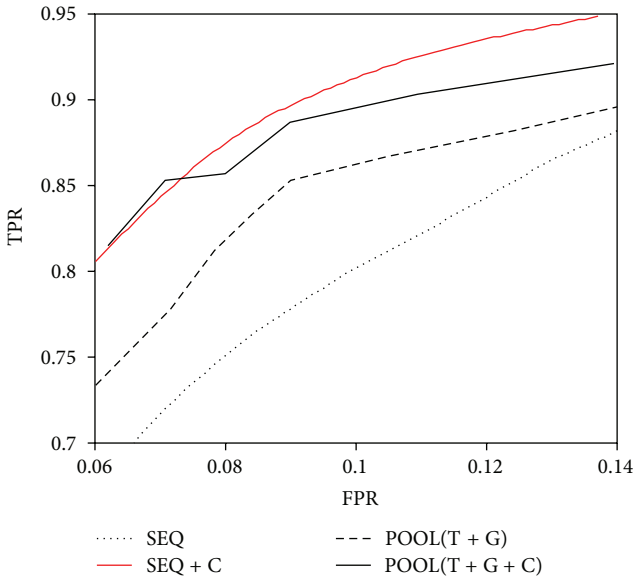


Figure 6: Comparison of ROC curves of SEQ and POOL on a dataset of 79 enzymes. Comparison of SEQ and POOL that combines all of its features, including POOL(T): residue electrostatic, POOL(G): structure geometry, dand POOL(C): sequence conservation. ROC curves of prediction using SEQ and SEQ combined with sequence conservation (SEQ + C) are both shown in the figure. The figure was remade from [6].

catalytic residue). In Figure 2(a), it is clear that SEQ correctly identifies all catalytic residues, which are located on the globally most rigid (small SEQ values) regions. For levansucrase, the three catalytic residues, Asp86, Asp247, and Glu342, are ranked 6, 5, and 2, respectively, by POOL. In Figure 2(b), the three catalytic residues locate in the most structurally rigid regions according to SEQ. It is also interesting that residues Glu340, Glu262, and Tyr411 are ranked 1, 3, 4, respectively, by POOL. These residues are located in relatively rigid regions in the SEQ profile. Table 3 lists the prediction rank of POOL and our prediction using SEQ ($n = 1$) for each catalytic residue in several example proteins. The rank of our prediction is based on the probability of a residue predicted to be catalytic residue based on the function provided by the LIBSVM software. The results show that our prediction results are in general better than or comparable to those of POOL in these examples.

3.5. Discussions on Related Prediction Methods. Here we discuss related catalytic residue prediction methods, including their features, datasets, and prediction performance. Petrova and Wu [34] used 24 features, including sequence-based features: amino acid type, sequence conservation, and structure-based and chemical features: shape of local structure, solvent accessible surface, structural flexibility, and hydrogen bonding. A dataset of 79 enzymes containing totally 23,664 residues and 254 catalytic residues was used for performance evaluation. Among these features, the seven best features were selected. The MCC of using different combinations of these features ranges from 0.52 to 0.74 and the sensitivities range from 0.88 to 0.89. To avoid the problems in SVM training due to the extremely unbalanced number of catalytic and noncatalytic residues, they used a similar strategy we used here for SVM training and predicting. The strategy is to build a subset that includes all catalytic residues and equal number of noncatalytic residues selected randomly. It is interesting to note that, without using the sequence conservation feature, the prediction MCC is only 0.52.

Table 3: Comparison of rank of catalytic residues for predictions using SEQ and POOL.

PDB ID and chain	Catalytic residue	Rank	
		SEQ ($n = 1$)[a]	POOL
1BWZ:A	C73	8	18
	H159	10	3
	E208	5	6
	C217	7	59
1OYG:A	D86	8	6
	D247	9	5
	E342	7	2
1A95:C	D88	1	1
	D89	2	2
	D92	7	8
	K115	6	27
1EC9:A	K205	3	10
	K207	2	4
	D313	8	3
	H339	7	1
	D366	5	12
1EHK:A	H233	2	6
	Y237	6	5
	H384	3	8
	F385	4	66
	H386	5	26
	R449	1	4
	R450	7	7

[a]The rank of prediction using SEQ is based on the probability of a residue predicted to be catalytic residue.

A more recent study by Cilia and Passerini [35] models spherical regions around target residues and extracts the properties of their content such as physicochemical properties, atomic density, flexibility, and presence of water molecules. They performed the prediction using SVM with these structural features and other sequence-based features: amino acid type and sequence conservation.

Amitai et al. [11] applied graph theory to catalytic residue prediction by converting protein structure to network in which the graph nodes are residues and graph edges are residues interactions. They found that catalytic sites have higher network closeness than noncatalytic residues. The features used in their prediction included the closeness feature, solvent accessible surface, and sequence conservation using a dataset of 178 enzymes.

Wie et al. [12] used a computational method called Theoretical Microscopic Titration Curves (THEMATICS), which computes theoretical electrostatic properties of residues based on structure information. They simply set a threshold to identify catalytic residues, that is, each residue was assigned a score calculated by THEMACTICS and residues having score greater than the threshold are predicted as catalytic residues. The dataset used contains 169 enzymes, including 594 annotated catalytic residues. The sensitivities using different thresholds range from 0.41 to 0.63. THEMATICS is then combined with other structure feature and is called POOL, with which we chose to compare our results. The reason we compared with POOL is that POOL is the most accurate structure-based prediction method. There are other methods having better prediction results combining complicated sequence features and structure features [35]. However, they usually do not provide prediction results only using structure features.

4. Conclusions

In this work, we calculated theoretical structural flexibility for catalytic residues and their sequence or structure neighboring residues. We found that catalytic residues are in general located in structurally less flexible context. We show that the theoretical structure flexibility (SEQ and STR) we used is better than B-factor for identification of catalytic residues. For a dataset of 760 enzymes of low pairwise sequence identity, the difference of SEQ distributions between catalytic and noncatalytic residues are more obvious than that of B-factor. The prediction results of SEQ are much better then those of B-factor. The MCC, sensitivity, and specificity of prediction are 0.74, 0.76, and 0.51, respectively, using SEQ combined with amino acid type information. The prediction results using SEQ are comparable to or better than those of other structure-based features. Most current prediction methods need homology information, for example, sequence conservation from PSI-Blast, and require the existence of sequence or structure similar proteins. SEQ and STR are calculated from single-protein structure and do not require any homology information. They may be further applied to the detection of enzyme function-related sites, like protein ligand binding site, metal binding site, or protein-protein interaction hotspot residues.

Acknowledgment

This research was supported by Grant (101-2218-E-320-001-) of National Science Council, Taiwan.

References

[1] H. M. Berman, J. Westbrook, Z. Feng et al., "The protein data bank," *Nucleic Acids Research*, vol. 28, no. 1, pp. 235–242, 2000.

[2] J. A. Capra and M. Singh, "Predicting functionally important residues from sequence conservation," *Bioinformatics*, vol. 23, no. 15, pp. 1875–1882, 2007.

[3] D. La, B. Sutch, and D. R. Livesay, "Predicting protein functional sites with phylogenetic motifs," *Proteins*, vol. 58, no. 2, pp. 309–320, 2005.

[4] M. Ota, K. Kinoshita, and K. Nishikawa, "Prediction of catalytic residues in enzymes based on known tertiary structure, stability profile, and sequence conservation," *Journal of Molecular Biology*, vol. 327, no. 5, pp. 1053–1064, 2003.

[5] B. Sterner, R. Singh, and B. Berger, "Predicting and annotating catalytic residues: an information theoretic approach," *Journal of Computational Biology*, vol. 14, no. 8, pp. 1058–1073, 2007.

[6] J. W. Torrance, G. J. Bartlett, C. T. Porter, and J. M. Thornton, "Using a library of structural templates to recognise catalytic sites and explore their evolution in homologous families," *Journal of Molecular Biology*, vol. 347, no. 3, pp. 565–581, 2005.

[7] N. Nagano, C. A. Orengo, and J. M. Thornton, "One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions," *Journal of Molecular Biology*, vol. 321, no. 5, pp. 741–765, 2002.

[8] A. C. Wallace, R. A. Laskowski, and J. M. Thornton, "Derivation of 3D coordinate templates for searching structural databases: application to Ser-His-Asp catalytic triads in the serine proteinases and lipases," *Protein Science*, vol. 5, no. 6, pp. 1001–1013, 1996.

[9] S. Sacquin-Mora, E. Laforet, and R. Lavery, "Locating the active sites of enzymes using mechanical properties," *Proteins*, vol. 67, no. 2, pp. 350–359, 2007.

[10] A. Ben-Shimon and M. Eisenstein, "Looking at enzymes from the inside out: the proximity of catalytic residues to the molecular centroid can be used for detection of active sites and enzyme-ligand interfaces," *Journal of Molecular Biology*, vol. 351, no. 2, pp. 309–326, 2005.

[11] G. Amitai, A. Shemesh, E. Sitbon et al., "Network analysis of protein structures identifies functional residues," *Journal of Molecular Biology*, vol. 344, no. 4, pp. 1135–1146, 2004.

[12] Y. Wei, J. Ko, L. F. Murga, and M. J. Ondrechen, "Selective prediction of interaction sites in protein structures with THEMATICS," *BMC Bioinformatics*, vol. 8, article 119, 2007.

[13] W. Tong, Y. Wei, L. F. Murga, M. J. Ondrechen, and R. J. Williams, "Partial Order Optimum Likelihood (POOL): maximum likelihood prediction of protein active site residues using 3D structure and sequence properties," *PLoS Computational Biology*, vol. 5, no. 1, Article ID e1000266, 2009.

[14] Y. T. Chien and S. W. Huang, "Accurate prediction of protein catalytic residues by side chain orientation and residue contact density," *PLoS ONE*, vol. 7, Article ID e47951, 2012.

[15] S. W. Huang, C. H. Shih, C. P. Lin, and J. K. Hwang, "Prediction of NMR order parameters in proteins using weighted protein contact-number model," *Theoretical Chemistry Accounts*, vol. 121, no. 3-4, pp. 197–200, 2008.

[16] C. P. Lin, S. W. Huang, Y. L. Lai et al., "Deriving protein dynamical properties from weighted protein contact number," *Proteins*, vol. 72, no. 3, pp. 929–935, 2008.

[17] Y. T. Chien and S. W. Huang, "Prediction of protein catalytic residues by local structural rigidity," in *Proceedings of the 6th International Conference on Complex, Intelligent and Software Intensive Systems (CISIS '12)*, pp. 592–596, Palermo, Italy, 2012.

[18] C. H. Q. Ding and I. Dubchak, "Multi-class protein fold recognition using support vector machines and neural networks," *Bioinformatics*, vol. 17, no. 4, pp. 349–358, 2001.

[19] C. S. Yu, J. Y. Wang, J. M. Yang, P. C. Lyu, C. J. Lin, and J. K. Hwang, "Fine-grained protein fold assignment by support vector machines using generalized npeptide coding schemes and jury voting from multiple-parameter sets," *Proteins*, vol. 50, no. 4, pp. 531–536, 2003.

[20] C. S. Yu, C. J. Lin, and J. K. Hwang, "Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions," *Protein Science*, vol. 13, no. 5, pp. 1402–1406, 2004.

[21] S. Hua and Z. Sun, "Support vector machine approach for protein subcellular localization prediction," *Bioinformatics*, vol. 17, no. 8, pp. 721–728, 2001.

[22] S. Hua and Z. Sun, "A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach," *Journal of Molecular Biology*, vol. 308, no. 2, pp. 397–407, 2001.

[23] H. Kim and H. Park, "Protein secondary structure prediction based on an improved support vector machines approach," *Protein Engineering*, vol. 16, no. 8, pp. 553–560, 2003.

[24] B. Rost and C. Sander, "Prediction of protein secondary structure at better than 70% accuracy," *Journal of Molecular Biology*, vol. 232, no. 2, pp. 584–599, 1993.

[25] Y. C. Chen and J. K. Hwang, "Prediction of disulfide connectivity from protein sequences," *Proteins*, vol. 61, no. 3, pp. 507–512, 2005.

[26] Y. C. Chen, Y. S. Lin, C. J. Lin, and J. K. Hwang, "Prediction of the bonding states of cysteines using the support vector machines based on multiple feature vectors and cysteine state sequences," *Proteins*, vol. 55, no. 4, pp. 1036–1042, 2004.

[27] S. W. Huang and J. K. Hwang, "Computation of conformational entropy from protein sequences using the machine-learning method—application to the study of the relationship between structural conservation and local structural stability," *Proteins*, vol. 59, no. 4, pp. 802–809, 2005.

[28] H. Kim and H. Park, "Prediction of protein relative solvent accessibility with support vector machines and long-range interaction 3D local descriptor," *Proteins*, vol. 54, no. 3, pp. 557–562, 2004.

[29] C. C. Chang and C. J. Lin, LIBSVM: a library for support vector machines, 2001, http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[30] Y. R. Tang, Z. Y. Sheng, Y. Z. Chen, and Z. Zhang, "An improved prediction of catalytic residues in enzyme structures," *Protein Engineering, Design and Selection*, vol. 21, no. 5, pp. 295–302, 2008.

[31] S. F. Altschul, T. L. Madden, A. A. Schäffer et al., "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, no. 17, pp. 3389–3402, 1997.

[32] C. T. Porter, G. J. Bartlett, and J. M. Thornton, "The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data," *Nucleic Acids Research*, vol. 32, pp. D129–D133, 2004.

[33] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers*, vol. 22, no. 12, pp. 2577–2637, 1983.

[34] N. V. Petrova and C. H. Wu, "Prediction of catalytic residues using Support Vector Machine with selected protein sequence and structural properties," *BMC Bioinformatics*, vol. 7, article 312, 2006.

[35] E. Cilia and A. Passerini, "Automatic prediction of catalytic residues by modeling residue structural neighborhood," *BMC Bioinformatics*, vol. 11, article 115, 2010.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 813912, 7 pages
http://dx.doi.org/10.1155/2013/813912

Research Article

Simpute: An Efficient Solution for Dense Genotypic Data

Yen-Jen Lin,[1] Chun-Tien Chang,[1] Chuan Yi Tang,[1,2] and Wen-Ping Hsieh[3]

[1] *Department of Computer Science, National Tsing Hua University, Hsinchu, Taiwan*
[2] *Department of Computer Science and Information Engineering, Providence University, Taichung, Taiwan*
[3] *Institute of Statistics, National Tsing Hua University, Hsinchu, Taiwan*

Correspondence should be addressed to Chuan Yi Tang; cytang@pu.edu.tw and Wen-Ping Hsieh; wphsieh@stat.nthu.edu.tw

Received 27 November 2012; Accepted 4 January 2013

Academic Editor: Hao-Teng Chang

Single nucleotide polymorphism (SNP) data derived from array-based technology or massive parallel sequencing are often flawed with missing data. Missing SNPs can bias the results of association analyses. To maximize information usage, imputation is often adopted to compensate for the missing data by filling in the most probable values. To better understand the available tools for this purpose, we compare the imputation performances among BEAGLE, IMPUTE, BIMBAM, SNPMStat, MACH, and PLINK with data generated by randomly masking the genotype data from the International HapMap Phase III project. In addition, we propose a new algorithm called simple imputation (Simpute) that benefits from the high resolution of the SNPs in the array platform. Simpute does not require any reference data. The best feature of Simpute is its computational efficiency with complexity of order $(mw + n)$, where n is the number of missing SNPs, w is the number of the positions of the missing SNPs, and m is the number of people considered. Simpute is suitable for regular screening of the large-scale SNP genotyping particularly when the sample size is large, and efficiency is a major concern in the analysis.

1. Background

A single nucleotide polymorphism (SNP) is a genetic variation at a single base-pair position. It is acquired and retained in the population. Most SNPs produce no observable difference between members of a species. These variations in the DNA can occur on both coding and noncoding sequences at a frequency of approximately 1 per 1000 base pairs [1, 2]. This leads to a rate of an estimated 11 million loci that can vary in approximately 0.1% of the population according to neutral theory of population genetics [3].

Studies concerning genetic association examine genetic traits shared among individuals in a population. SNPs have an important role in these studies because they record the history of recombination and are sufficiently dense to form linkage disequilibrium (LD) in nearly all functional genes. However, it is common for data to be missing on the various genotyping platforms. Even for array technology, the rate of missing data can be as high as 0.53% [4]. This is approximately 5300 loci for every million SNPs designed on the arrays. Assuming a random missing mechanism exists, if any locus in a sample is removed, the missing rate can become as high as $1 - (1 - 0.0053)^n$ in an association study of n samples.

Because it is often not financially viable to regenotype the missing data, imputation is used to fill in the missing SNP values, and to maintain low costs. Imputation can be as simple as selecting at random a genotype that already exists in the data or by using a major allele. However, such naive methods normally result in high error rates [4]. Certain other methods are based on haplotypes, which are sets of SNPs that are associated on one chromosome pair. These methods include the Hidden Markov Model (HMM), Markov chain (MC), maximum-likelihood, and neural network. Because a multitude of methodologies exists that can be employed to impute a haplotype, a range of imputation software, consequently, also exists. Examples of imputation software include IMPUTE [5], MaCH [6], SNPMSTAT [7], fastPhase [8], and BEAGLE [9].

Both the IMPUTE and BEAGLE software use the HMM. The HMM is a statistical tool for modelling generative

sequences, which are characterised by the use of an underlying process to generate an observable sequence. In the HMM these underlying processes are represented by states, which are considered to be unobserved or hidden. The hidden state used is a pair of haplotypes observed in reference samples from the HapMap project. The observed data are the individual genotypes at the corresponding loci. IMPUTE considers mutation and recombination in its HMM model; this requires additional information from CHIAMO [10] and HAPGEN [6, 10, 11] to determine the probability of the genotypes estimated from the arrays and the predicted haplotypes. MaCH uses a Markov chain-based approach using samples from HapMap as references. Long missing segments are compensated for in MaCH by using haplotypes from the reference samples.

Alternative imputation software and methodologies include SNPMSTAT and FFNN. SNPMSTAT uses a maximum-likelihood framework on the genotype data. It uses HapMap data or other similar data sets to construct the most-likely haplotypes to occur for a missing SNP value. The feed-forward neural networks (FFNNs) proposed by Sun and Kardia [12] were reported to perform well by using a Bayesian criterion to select the predictors. They claimed that the performance is better than that of fastPHASE [8] and the LD-based method, which is used by HelixTree [10].

In this paper, we propose an algorithm based on observed genotypes and the LD at three neighbouring SNPs, including the SNP under consideration, to impute the missing SNPs, and to reduce the error rate for estimation. This algorithm only considers the two neighbouring SNPs and uses the haplotype information, which is a direct consequence of LD. Jung et al. used the same level of information in their proposed method [4], which phased genotypes by the partition-ligation expectation maximization (PLEM) [11] to impute the missing SNPs. We compare the results using SNPs from the same chromosomal regions in Jung's study and demonstrate the better performance of our algorithm. We also compare the general-purpose methods including BIMBAM v0.99, BEAGLE v3.0.3, IMPUTE v0.5.0, MARCH v1.0.16, PLINK, and SNPMSTAT v3.1. Because Simpute provides the best power at highly linked loci, we compare it to the best method using SNPs with strong LD. We demonstrate that Simpute is a promising tool to provide efficient computation when it comes to the age of massive parallel sequencing.

2. Methods

SNPs could be bi-, tri-, or tetraallelic polymorphisms by definition, but triallelic and tetraallelic SNPs rarely exist in the human population. SNPs are usually considered biallelic, and three genotypes are possible for each SNP locus. They are coded as 0 (homozygous for the wild type), 2 (heterozygote), and 1 (homozygous for mutants) in this study.

Two neighbouring SNP loci of the missing target are considered in the Simpute method. Haplotypes formed by the consecutive pair of loci are constructed and the estimated haplotype probabilities are combined with the LD information from either side of the missing SNP to predict the missing SNP genotype.

2.1. Estimate the Population Proportion of Haplotypes. We first considered genotypes at two loci. The counts of all genotype combinations are summarized in Table 3.

In Table 3, there are nine genotype combinations. The haplotypes for eight of them can be clearly resolved, while those of the $N_{1,1}$ could be either ab/AB or aB/Ab. The proportion of the four haplotypes can be estimated as follows:

$$\begin{aligned} p(\text{ab}) &= \frac{2 \times N_{0,0} + N_{0,1} + N_{1,0} + X_1 \times N_{1,1}}{2 \times N_{P,Q}}, \\ p(\text{aB}) &= \frac{2 \times N_{0,2} + N_{0,1} + N_{1,2} + X_2 \times N_{1,1}}{2 \times N_{P,Q}}, \\ p(\text{Ab}) &= \frac{2 \times N_{2,0} + N_{2,1} + N_{1,0} + X_2 \times N_{1,1}}{2 \times N_{P,Q}}, \\ p(\text{AB}) &= \frac{2 \times N_{2,2} + N_{2,1} + N_{1,2} + X_1 \times N_{1,1}}{2 \times N_{P,Q},}, \end{aligned} \tag{1}$$

where X_1 is the proportion of the phase ab/AB with the observed genotype aAbB, and X_2 is the proportion of the phase aB/Ab.

The initial values for X_1 and X_2 are set to 0.5, and they are iteratively updated to get a more probable estimate. The updating step is

$$\begin{aligned} X_1 &= \frac{p(\text{ab}) \times p(\text{AB})}{p(\text{ab}) \times p(\text{AB}) + p(\text{aB}) \times p(\text{Ab})}, \\ X_2 &= \frac{p(\text{aB}) \times p(\text{Ab})}{p(\text{ab}) \times p(\text{AB}) + p(\text{aB}) \times p(\text{Ab})}. \end{aligned} \tag{2}$$

The estimated X_1 and X_2 are then used to calculate the p(ab), p(aB), p(Ab), and p(AB) in (1). The 10 iterations will stop for either X_1 or X_2. According to (1) and (2), X_1 or X_2 is a cubic function, solved by the cubic formula. Here we use the iterative method to solve X_1 and X_2. The initial value of both is set to 0.5, where the two phases have the same probability (Table 4).

2.2. Linkage Disequilibrium Measurement. We impute the missing genotypes using the LD information and the haplotype probabilities calculated in the previous section. If the LD value between two SNP sites is high, then the two SNPs are close to each other, and there are relatively few recombination events between them. Some measurements are commonly used to evaluate the extent of LD between a pair of SNP sites. Two important pairwise measures of LD are r^2 and $|D'|$ [13–15]. Their range is from 0 to 1, but their interpretation is slightly different. When $|D'|$ is equal to 1, r^2 can be small. For example, when p(ab) = 0.9, p(aB) = 0.1, p(Ab) = 0.1, and p(AB) = 0, $|D'|$ is equal to 1, the r^2 value is 0.012. In this paper, r^2 is derived from the input samples. The $|D'|$ and r^2 can be computed as follows.

The difference between the observed and the expected probability of two loci is measured. The disequilibrium coefficient D is expressed as

$$D = p(\text{ab}) - p(\text{a}\cdot) \times p(\cdot\text{b}). \tag{3}$$

The normalized disequilibrium coefficient is defined as $D' = D/|D|_{\max}$ according the study of Pritchard and Przeworski [14], where

$$D_{\max} = \begin{cases} \min\left(p\left(\mathrm{a}\cdot\right) \times p\left(\cdot\mathrm{B}\right), p\left(\mathrm{A}\cdot\right) \times p\left(\cdot\mathrm{b}\right)\right), & \text{if } D \geq 0 \\ \min\left(p\left(\mathrm{a}\cdot\right) \times p\left(\cdot\mathrm{b}\right), p\left(\mathrm{A}\cdot\right) \times p\left(\cdot\mathrm{B}\right)\right), & \text{if } D < 0. \end{cases} \tag{4}$$

The range of the normalized disequilibrium coefficient D' is $[-1, 1]$. D' can be 1 while the P value is not significant. That is, when D' is equal to 1, there can still be no association. Hence, we adopt another popular measurement r^2, where

$$r^2 = \frac{D^2}{p\left(\mathrm{a}\cdot\right) \times p\left(\cdot\mathrm{b}\right) \times p\left(\mathrm{A}\cdot\right) \times p\left(\cdot\mathrm{B}\right)}. \tag{5}$$

The r^2 value between the sites P and Q is denoted as $r^2_{P,Q}$.

2.3. Imputation Algorithm. Consider three SNP sites P, Q, and R that are in consecutive order. The imputation procedure is as follows.

(1) Use the samples with no missing data at P, Q, and R to calculate the pairwise r^2 at loci P, Q, and R. If the r^2 equals zero, it will be set to a minimum value of 10^{-5} to facilitate the following computation.

(2) Because most haplotypes consisting of three loci are rare in the population, and the population proportion cannot be correctly estimated with the limited samples in most studies, we approximate it with the product of haplotype proportion for the three pairs of loci and put the LD measured between the two loci as the weights. The probability for haplotype $h_1h_2h_3$ is approximated as

$$\begin{aligned} P_{P,Q,R}\left(h_1h_2h_3\right) = {} & P_{P,Q}\left(h_1h_2\right) \times r^2_{P,Q} \\ & \times P_{Q,R}\left(h_2h_3\right) \times r^2_{Q,R} \times P_{P,R}\left(h_1h_3\right) \times r^2_{P,R}, \end{aligned} \tag{6}$$

where $P_{P,Q}(h_1h_2)$, $P_{Q,R}(h_2h_3)$, and $P_{P,R}(h_1h_3)$ are the probabilities of haplotype h_1h_2 at loci P, Q, haplotype h_2h_3 at loci Q, R, and haplotype h_1h_3 at loci P, R. These probabilities were generated by (1).

(3) Calculate the weighting score of genotype $\otimes\oplus$ at each pair of loci:

$$W_{(\otimes,\oplus)} = 1 - \left|\frac{N_{\otimes,\cdot} \times N_{\cdot,\oplus}}{N \times N} - \frac{N_{\otimes,\oplus}}{N}\right|, \tag{7}$$

where $\otimes$ and $\oplus$ are the genotypes at the first and the second locus in each pair. If the $W_{(\otimes,\oplus)}$ equals zero, it will be set to a minimum value of 10^{-5} to facilitate the following computation.

(4) Calculate the haplotype pair score

$$\begin{aligned} \text{score} = {} & \left(P_{P,Q,R}\left(h_1h_2h_3\right) + P_{P,Q,R}\left(h'_1h'_2h'_3\right)\right) \\ & \times \frac{N^{P,Q}_{\otimes,\oplus} \times N^{P,R}_{\otimes,\circ} \times N^{Q,R}_{\oplus,\circ}}{N^{P,Q}_{\otimes,\cdot} \times N^{P,R}_{\cdot,\circ} \times N^{Q,R}_{\oplus,\cdot}} \times W_{(\otimes,\oplus)} \times W_{(\otimes,\circ)} \times W_{(\oplus,\circ)}, \end{aligned} \tag{8}$$

where the probabilities of the haplotype pair $P_{P,Q,R}(h_1h_2h_3)$ and $P_{P,Q,R}(h'_1h'_2h'_3)$ are calculated by (6), and $\otimes$, $\oplus$, $\circ$ represent the same genotypes $(h_1h'_1)$, $(h_2h'_2)$, and $(h_3h'_3)$ at locus P, Q, and R, respectively.

(5) Choose from all legitimate haplotype pairs that maximize the score in (8).

The algorithm also considers the situation when consecutive SNPs are missing. In that case, the two neighbouring loci P and R of the missing locus Q can represent the adjacent two loci on the same side of the Q. For example, when there is a long stretch of missing genotypes from SNP 1 to 4 in a specific sample, we can first impute locus 4 with information from locus 5 and 6 and then sequentially fill in all the missing ones.

2.4. Time Complexity. Our algorithm requires the computation complexity at the order of $O(mw + n)$ where n is the number of missing SNPs, w is the number of the SNP loci with at least one missing entry, and m is the number of individual with at least one locus missing. Each sample requires the order of $O(1)$ to count each of the 9 genotype and the order of $O(mw)$ for steps 1 and 2. Hence, the total complexity of the algorithm is $O(mw + n)$.

3. Data Description

In this paper, we used two data sets to compare imputation performance. All data sets are based on the individuals included in the HapMap project [16].

3.1. SNP-Dense Region on Chromosome 22. The first data set was the testing region adopted from Jung et al. They identified a region with dense SNP distribution and demonstrated their performance with only six SNPs, as annotated in HapMap Phase II, release 22. Those SNPs are rs2213329, rs2227029, rs9610029, rs2213331, rs9619447 and rs743726, and are located from positions 33227611 to 33233156 of chromosome 22. This region was selected for its strong linkage of $|D'| > 0.7$. We used the SNP data of 270 people from HapMap to generate the testing data. The data were randomly selected with missing rates of 5%, 10%, 15%, and 20% from the total of $270 \times 6 = 1620$ SNPs. We adopted the settings of the missing rates of Jung et al. for comparison purposes. This random procedure was repeated 100 times, and the average error rates were obtained. A more realistic comparison is demonstrated with the other set of random missing studies described in the following section.

3.2. Random Missing SNPs from the HapMap Phase III on Chromosome 21. We used samples of HapMap phase III as our testing data. Because some of the software we compared required reference data, we provided samples of HapMap Phase II release 22 as the reference samples; those samples were, thus, excluded in our testing set. SNP loci that are tri-alleic or tetraallelic were excluded in the comparison; Tables 1 and 2 show the proportion of this type of loci in the reference samples (HapMap Phase II release 22) and testing samples (HapMap Phase III specific samples), respectively.

Table 1: The nonbiallelic loci proportion in the HapMap phase II release 22.

Population	Individuals	SNPs	Nonbiallelic
CEU	90	48217	1.69%
JPT + CHB	90	50053	1.81%
YRI	90	48541	1.60%

Table 2: The non-bi-allelic loci proportion in the HapMap phase III.

Population	Individuals	SNPs	Non-bi-allelic
CEU	80	19250	0.39%
JPT + CHB	77	17286	0.21%
YRI	80	20198	0.21%

Table 3: A 3 × 3 contingency table for the genotypes at two consecutive loci. A and a are the two alleles in locus 1 while B and b are the two alleles in locus 2.

	0 (bb)	1 (bB)	2 (BB)	Total
0 (aa)	$N_{0,0}$	$N_{0,1}$	$N_{0,2}$	$N_{0,\cdot}$
1 (aA)	$N_{1,0}$	$N_{1,1}$	$N_{1,2}$	$N_{1,\cdot}$
2 (AA)	$N_{2,0}$	$N_{2,1}$	$N_{2,2}$	$N_{2,\cdot}$
Total	$N_{\cdot,0}$	$N_{\cdot,1}$	$N_{\cdot,2}$	$N_{P,Q}$

The proportion is low and is not crucial for the conclusion. We conducted the experiment on the smallest chromosome to enable easier computation of the less efficient algorithms in the comparison. The results are reported separately for the different ethnic groups because certain interesting differences were observed.

We generated three sets of testing data from the HapMap Phase III specific samples. The first set was derived by randomly masking the genotypes on chromosome 21, called the *complete set*. Because the error rate of genotype calling is less than 1% [17], the missing rates were 0.1%, 0.5%, 1%, and 5%. Ten randomly missing testing data sets were generated for comparison, and the accuracy was calculated as the average of the 10 repeats. The software used to compare the data set included BIMBAM v0.99, BEAGLE v3.0.3, IMPUTE v0.5.0, MARCH v1.0.16, PLINK, and SNPMSTAT v3.1 and used the system Linux kernel version 2.6 on AMD 64 platform.

Our second test data consisted of numerous regions of only three SNPs on chromosome 21, called the *short input*. At most, two of the three SNPs were permitted to be missing in our random sampling process. The error rates are reported, with the averages of 25 repeats of the random missing procedure for missing rates, as 0.1%, 0.5%, 1%, and 5%.

The algorithm we proposed adopted minimum information to complete the missing gaps, and, hence, it is not designed for all purposes. We show that its performance at the highly linked regions is no worse than the best method previously mentioned. The third set of test data consists of missing SNPs with strong linkage ($r^2 > 0.9$) to either one of their adjacent SNPs, called *high LD*. The advantage acquired at the highly linked regions is the most important aspect of Simpute and is why Simpute is the most helpful program in global genome sequencing projects. The error rates are reported from the average of 100 repeats of the random missing procedure for the missing rates at 0.1%, 0.5%, 1%, and 5%.

Table 4: Notation for the haplotype probabilities at the two loci.

Locus $P \setminus Q$	b	B	Total
a	p (ab)	p (aB)	p (a·)
A	p (Ab)	p (AB)	p (A·)
Total	p (·b)	p (·B)	1

Table 5: Error rates* for Simpute, BEAGLE, and Jung's method with random missing study on the six SNPs of chromosome 22.

Missing rate/method	Simpute	BEAGLE	Jung's method
5%	1.358%	1.7531%	16.59%
10%	1.8944%	2.1429%	17.82%
15%	3.0207%	3.4132%	20.25%
20%	4.4472%	4.4907%	20.07%

*Error rates = number of error imputed entries/number of missing entries ∗100%.

4. Results

We used samples from HapMap Phase II release 22 as the reference data set, which is required by BEAGLE, BIMBAM, MACH, SNPMStat, IMPUTE, and plink. Because of the intractable computation load of SNPMStat and IMPUTE, we divided the chromosome into segment of 10,000 SNPs for the inputs. Because SNPMStat requires substantial CPU time, only three repeats were performed to derive the average accuracy. All the other programs used 10 repeats to obtain the averages. The results from the *complete set* are shown in Figure 1. BEAGLE gives the best overall accuracy and is also the fastest on our benchmark platform CentOS 5.3 under the VNWare ESX 4i in Figure 2. The following comparisons only address the differences between Simpute and BEAGLE.

The results from the SNP-dense region of chromosome 22 in Jung's study are shown in Table 5. The error rates from the Jung et al. study are copied directly from their report because we did not implement their algorithm. It appears that Simpute has a strong advantage in the SNP-dense regions. Although BEAGLE used the same HapMap samples as reference samples and used all six SNPs together in their complicated algorithms, it still has slightly higher error rates, and the contrast is strong at the lower missing rates.

To understand the relation between the information fed into each method and the power each method gains, we first assessed the sets of three SNPs on chromosome 21. This provided limited information, and the error rates for Simpute and BEAGLE are shown in Table 6. The missing rates were set as 0.1%, 0.5%, 1%, and 5% to better match the actual applications. Because the data are artificial and require repeated initiation processes for BEAGLE to process all the short regions, extensive computation time is required for BEAGLE to process all the data. Hence, comparing the computation time is not feasible, and it is difficult to run the entire set of simulations on all three ethnic groups. We only reported the results for Group CEU with 25 repeats of the random missing

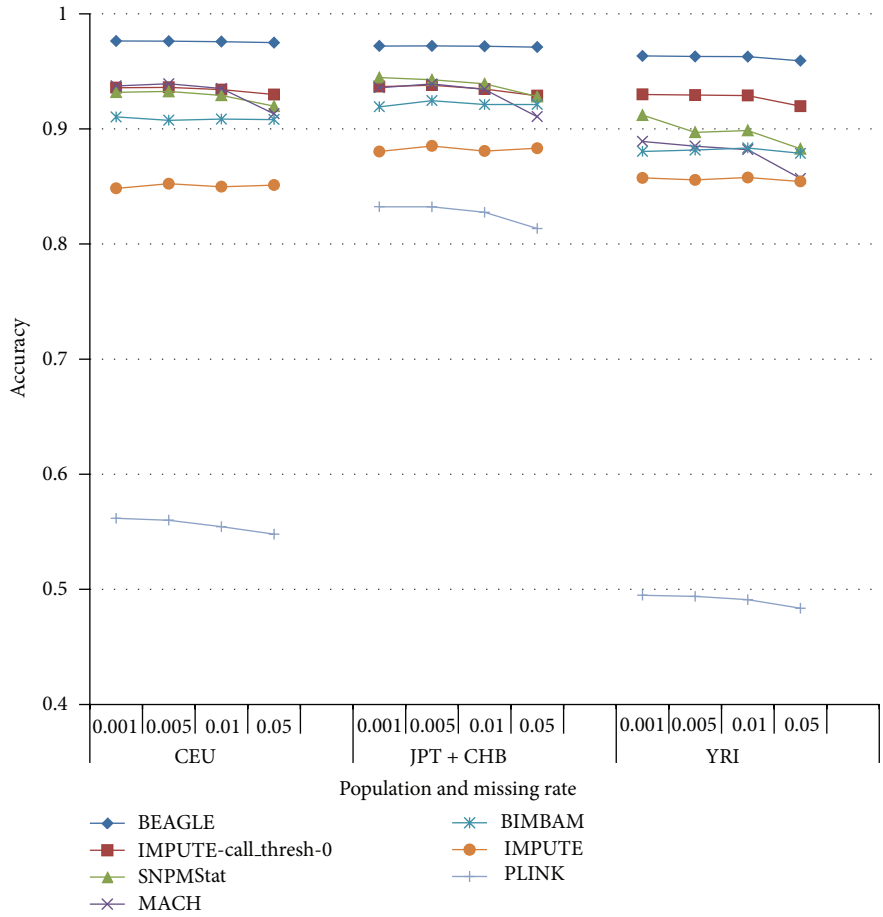


Figure 1: Imputation accuracy compared across BEAGLE, IMPUTE, BIMBAM, SNPMStat, MACH, and plink using the *complete set*. The curve with IMPUTE-call_thresh-0 stands for the best setting (call thresh = 0) we found for Impute rather than the default setting. Accuracy = number of correctly imputed entries/number of missing entries ∗100%.

procedure, as shown in Table 6. The error rate of Simpute is approximately the same as that of BEAGLE and matches our expectations.

Tables 7, 8, and 9 show the evaluation of Simpute and BEAGLE using the *high LD* testing data on chromosome 21. The default setting of BEAGLE used the same 270 people from HapMap Phase II as the reference data. In contrast, Simpute used the two neighboring SNPs of the missing one. The error counts are the averages of 100 repeats of the random missing procedure. BEAGLE performed better than Simpute but the difference is negligible when the missing rate is low. In addition, BEAGLE requires substantially more processing time.

Table 6: The error rates* for random missing SNPs of short input at $r^2 \geq 0.9$ from the HapMap phase III on chromosome 21 of short input for the CEU.

Method/ missing rate	Simpute	BEAGLE
0.1%	37.136/483 (7.69%)	38.09/483 (7.89%)
0.5%	188/2412 (7.79%)	183.6364/2412 (7.61%)
1%	378.333/4823 (7.84%)	376.762/4823 (7.81%)
5%	1913.632/24111 (7.94%)	1892.053/24111 (7.84%)

*Error rates = number of error imputed entries/number of missing entries ∗100%.

5. Conclusion and Discussion

In this study we developed a simple strategy to impute missing genotypes for SNPs that have a high resolution. Our method requires only two neighbouring loci of a missing SNP. Furthermore, we show in our study that for highly linked loci, our algorithm has comparable performance to BEAGLE, a system that incorporates data from various sources of information, as has been suggested in recent studies. These sources of information include reference samples and long-range LD.

The algorithm we introduced in our study has a complexity of $O(mw + n)$, where n is the number of missing SNPs, w is the number of the positions of the missing SNPs, and m is the sample size. Because of the design of our algorithm, and the reduction of the prerequisite input incorporated into the imputation algorithm, we were able to significantly reduce the computation time.

Table 7: Error rates* and computation time for random missing SNPs of high LD for the CEU samples.

Method/missing rate	Simpute		BEAGLE	
	Error rate	Running time (sec)	Error rate	Running time (sec)
0.1%	5.52/483 (1.14%)	12.88	4.49/483 (0.93%)	164.17
0.5%	27.94/2412 (1.16%)	13.09	21.01/2412 (0.87%)	164.82
1%	57.22/4823 (1.19%)	14.07	44.07/4823 (0.91%)	168.47
5%	321.9/24111 (1.33%)	18.24	224.65/24111 (0.974%)	168.69

*Error rates = number of error imputed entries/number of missing entries ∗100%.

Table 8: Error rates* and computation time for random missing SNPs of high LD for the CHB + JPT samples.

Method/missing rate	Simpute		BEAGLE	
	Error rate	Running time (sec)	Error rate	Running time (sec)
0.1%	5.15/493 (1.04%)	10.90	4.64/493 (0.94%)	138.40
0.5%	27.29/2463 (1.10%)	11.08	24/2463 (0.97%)	139.79
1%	55.07/4925 (1.11%)	11.77	47.69/4925 (0.97%)	138.96
5%	322.38/24622 (1.31%)	16.113	242.26/24622 (0.98%)	140.96

*Error rates = number of error imputed entries/number of missing entries ∗100%.

Table 9: Error rates* and computation time for random missing SNPs of high LD for the YRI samples.

Method/missing rate	Simpute		BEAGLE	
	Error rate	Running time (sec)	Error rate	Running time (sec)
0.1%	2.57/271 (0.95%)	12.42	2.23/271 (0.82%)	187.80
0.5%	13.54/1353 (1.00%)	12.925	11.2/1353 (0.83%)	188.41
1%	27.19/2705 (1.00%)	13.08	22.89/2705 (0.85%)	187.45
5%	161.02/13525 (1.19%)	15.921	119.94/13525 (0.887%)	191.29

*Error rates = number of error imputed entries/number of missing entries ∗100%.

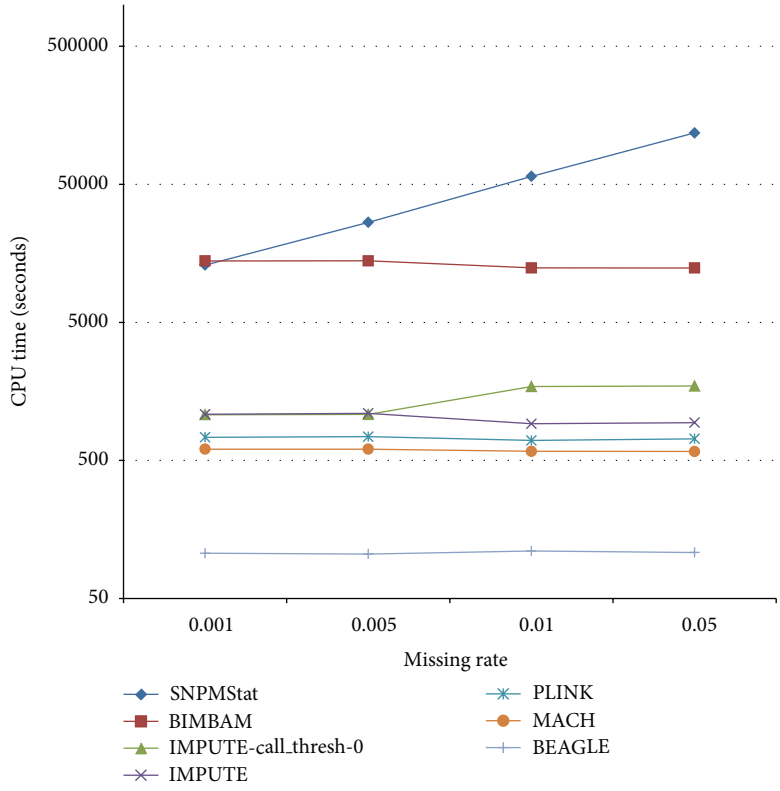


Figure 2: CPU Time.

Although Simpute is unable to outperform most software for general purposes, it has shown its potential for specific purposes. With the current trend of mass parallel-sequencing technologies, SNPs will soon be discovered with ease, without requiring the use of predefined positions for their detection. Furthermore, the availability of samples will accumulate in the following few years. Thus, it is expected that most SNPs will be highly linked in samples of moderate size.

Simpute has a strong advantage over more complicated algorithms that use high LD regions. Moreover, it demonstrates a distinct advantage in efficiency when handling large data sets. This efficiency is of great benefit to genome centers, which have increasing demands in the number of personal genomes that must be sequenced and analyzed through a real-time system.

Availability

Simpute is available from the following website: http://www.cs.nthu.edu.tw/~dr928307/Simpute.htm. We provide an integrated interface to run all of these softwares. It can be downloaded at http://kitty.2y.idv.tw/~tcs/ASHG2009/ and performed under Linux kernel 2.6 amd64 platform.

Authors' Contribution

Y.-J. Lin and C. T. Chang contributed equally to this work.

References

[1] J. I. Bell, "Single nucleotide polymorphisms and disease gene mapping," *Arthritis Research*, vol. 4, supplement 3, pp. S273–S278, 2002.

[2] R. Sachidanandam, D. Weissman, S. C. Schmidt et al., "A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms," *Nature*, vol. 409, no. 6822, pp. 928–933, 2001.

[3] L. Kruglyak and D. A. Nickerson, "Variation is the spice of life," *Nature Genetics*, vol. 27, no. 3, pp. 234–236, 2001.

[4] H. Y. Jung, Y. J. Park, Y. J. Kim, J. S. Park, K. Kimm, and I. Koh, "New methods for imputation of missing genotype using linkage disequilibrium and haplotype information," *Information Sciences*, vol. 177, no. 3, pp. 804–814, 2007.

[5] J. Marchini, B. Howie, S. Myers, G. McVean, and P. Donnelly, "A new multipoint method for genome-wide association studies by imputation of genotypes," *Nature Genetics*, vol. 39, no. 7, pp. 906–913, 2007.

[6] Y. Li, C. J. Willer, J. Ding, P. Scheet, and G. R. Abecasis, "MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes," *Genetic Epidemiology*, vol. 34, no. 8, pp. 816–834, 2010.

[7] D. Y. Lin, Y. Hu, and B. E. Huang, "Simple and efficient analysis of disease association with missing genotype data," *American Journal of Human Genetics*, vol. 82, no. 2, pp. 444–452, 2008.

[8] P. Scheet and M. Stephens, "A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase," *American Journal of Human Genetics*, vol. 78, no. 4, pp. 629–644, 2006.

[9] B. L. Browning and S. R. Browning, "A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals," *American Journal of Human Genetics*, vol. 84, no. 2, pp. 210–223, 2008.

[10] M. N. Chiano and D. G. Clayton, "Fine genetic mapping using haplotype analysis and the missing data problem," *Annals of Human Genetics*, vol. 62, no. 1, pp. 55–60, 1998.

[11] Z. S. Qin, T. Niu, and J. S. Liu, "Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms," *American Journal of Human Genetics*, vol. 71, no. 5, pp. 1242–1247, 2002.

[12] Y. V. Sun and S. L. R. Kardia, "Imputing missing genotypic data of single-nucleotide polymorphisms using neural networks," *European Journal of Human Genetics*, vol. 16, no. 4, pp. 487–495, 2008.

[13] B. Devlin and N. Risch, "A comparison of linkage disequilibrium measures for fine-scale mapping," *Genomics*, vol. 29, no. 2, pp. 311–322, 1995.

[14] J. K. Pritchard and M. Przeworski, "Linkage disequilibrium in humans: models and data," *American Journal of Human Genetics*, vol. 69, no. 1, pp. 1–14, 2001.

[15] R. C. Lewontin, "Interaction of selection and linkage. I. General considerations; heterotic models," *Genetics*, vol. 49, no. 1, pp. 49–67, 1964.

[16] K. A. Frazer, D. G. Ballinger, D. R. Cox et al., "A second generation human haplotype map of over 3.1 million SNPs," *Nature*, vol. 449, no. 7164, pp. 851–861, 2007.

[17] M. J. Huentelman, D. W. Craig, A. D. Shieh et al., "SNiPer: improved SNP genotype calling for Affymetrix 10K GeneChip microarray data," *BMC Genomics*, vol. 6, article 149, 2005.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 170398, 12 pages
http://dx.doi.org/10.1155/2013/170398

Research Article

In Silico Prediction and *In Vitro* Characterization of Multifunctional Human RNase3

Pei-Chun Lien,[1] Ping-Hsueh Kuo,[1] Chien-Jung Chen,[1] Hsiu-Hui Chang,[1] Shun-lung Fang,[1] Wei-Shuo Wu,[2] Yiu-Kay Lai,[2] Tun-Wen Pai,[3] and Margaret Dah-Tsyr Chang[1,4]

[1] *Institute of Molecular and Cellular Biology, National Tsing Hua University, No. 101, Section 2, Kuang Fu Road, Hsinchu 30013, Taiwan*
[2] *Institute of Biotechnology, National Tsing Hua University, No. 101, Section 2, Kuang Fu Road, Hsinchu 30013, Taiwan*
[3] *Department of Computer Science and Engineering, National Taiwan Ocean University, 2 Pei Ning Road, Keelung 20224, Taiwan*
[4] *Department of Medical Science, National Tsing Hua University, No. 101, Section 2, Kuang Fu Road, Hsinchu 30013, Taiwan*

Correspondence should be addressed to Margaret Dah-Tsyr Chang; dtchang@life.nthu.edu.tw

Received 31 October 2012; Accepted 2 December 2012

Academic Editor: Hao-Teng Chang

Human ribonucleases A (hRNaseA) superfamily consists of thirteen members with high-structure similarities but exhibits divergent physiological functions other than RNase activity. Evolution of hRNaseA superfamily has gained novel functions which may be preserved in a unique region or domain to account for additional molecular interactions. hRNase3 has multiple functions including ribonucleolytic, heparan sulfate (HS) binding, cellular binding, endocytic, lipid destabilization, cytotoxic, and antimicrobial activities. In this study, three putative multifunctional regions, ^{34}RWRCK^{38} (HBR1), ^{75}RSRFR^{79} (HBR2), and ^{101}RPGRR^{105} (HBR3), of hRNase3 have been identified employing *in silico* sequence analysis and validated employing *in vitro* activity assays. A heparin binding peptide containing HBR1 is characterized to act as a key element associated with HS binding, cellular binding, and lipid binding activities. In this study, we provide novel insights to identify functional regions of hRNase3 that may have implications for all hRNaseA superfamily members.

1. Introduction

Human ribonuclease A (hRNaseA) family members are encoded by unique genes located on human chromosome 14 [1]. The hRNaseA family is vertebrate cationic protein sharing conserved tertiary structure and specific enzymatic sites for RNase activity. It is in general considered to comprise eight members: RNase1 (pancreatic RNase), RNase2 (eosinophil derived neurotoxin/EDN), RNase3 (eosinophil cationic protein/ECP), RNase4, RNase5 (angiogenin), RNase6, RNase7 (skin-derived RNase), and RNase8 (divergent paralog of RNase7) [2]. Analysis of human genome sequence has revealed the existence of five additional RNases named as RNases9–13, although they appear to lose enzymatic activity [3]. All hRNaseA family members encode relatively small polypeptides of 14 to 16 kDa containing signal peptides of 20 to 28 amino acids for protein secretion. Mature hRNaseA members contain 6 to 8 cysteine residues that are crucial to hold the overall tertiary structure [4]. They possess an invariant catalytic triad including two histidines (one near the *N* terminus, and the other near the *C* terminus) and one lysine located within a conserved signature motif (CKXXNTF) [5]. These RNaseAs are catalytically active to various degrees against standard polymeric RNA substrates [6]. Interestingly, their host defense functions including cytotoxic [7, 8], helminthotoxic [9, 10], antibacterial [11, 12], and antiviral [5, 13] activities have also been reported. However, the mechanisms of noncatalytic functions of some hRNaseA members, especially the ones with low RNase activities, are poorly understood.

hRNase3 is found within the secondary granules of eosinophils and serves as a clinical asthma marker [14]. It is a multiple functional protein as the *N*-terminal domain[1–45]

possesses antipathogenic activities such as antibacterial, antihelminthic, and antiviral competencies [15–17]. In terms of key amino acids involving specific functions, Trp^{35} of hRNase3 interacts with cell membrane to form transmembrane pores in an artificial lipid bilayer, suggesting that lipid destruction is a crucial step in bactericidal activity [18]. In addition, Arg^{1}, Trp^{10}, Gln^{14}, Lys^{38}, and Gln^{40} located in antipathogenic domain are identified to bind to lipopolysaccharide (LPS) and peptidoglycan with high affinity, which may also be important for its bactericidal activity [19, 20]. Moreover, hRNase3 possesses cytotoxic activity against various mammalian cell lines including those derived from blood and epidermis [17], and Arg^{97} is the key residue for its cytotoxicity [21]. It is also highly associated with host inflammatory response and thus involved in tumor microenvironment to exercise its antitumor response [22–24]. Interestingly, hRNase3 reduces the infectivity of human respiratory syncytial virus in an RNase activity-dependent manner [25]. Treatment of bronchial epithelial cells with hRNase3 induces production of tumor necrosis factor alpha (TNF-α) and triggers apoptosis via a caspase-8-dependent pathway [26]. Furthermore, hRNase3 is reported to bind to a class of cell surface receptors termed as heparan sulfate proteoglycans [27] and thereby internalizes target cells through macropinocytosis [28]. Taken together, basic and aromatic residues, especially Arg and Trp, are considered to play important roles in enzymatic and other biological functions of hRNase3.

In this study, in combination with *in silico* analyses employing Reinforced Merging for Unique Segments (ReMUS) system, we have identified three heparin binding regions (HBRs) in hRNase3. We focused on their roles in heparin and cellular binding and endocytic and cytotoxic activities employing *in vitro* functional analyses. Our results showed that HBR1 (^{34}RWRCK^{38}) is crucial for enzymatic RNase function and serves as a major heparin binding site for endocytosis, HBR2 (^{73}RSRFR^{77}) contributes toward cell binding and endocytic activities, and HBR3 (^{101}RPGRR^{105}) plays a critical role in cytotoxicity. In addition, a noncytotoxic HBR1-derived peptide was characterized to bind to negatively charged molecules including glycosaminoglycans (GAGs) and lipids on cell surface. In summary, we have identified multifunctional regions of hRNase3, which may provide novel insights to implicate for all hRNaseA superfamily members.

2. Materials and Methods

2.1. In Silico Analysis. Unique peptides of query proteins, 13 hRNaseA family members, were identified employing Reinforced Merging for Unique Segments ReMUS system (ReMUS) (http://140.121.196.30/remus.asp) [42]. The system adopted a bottom-up strategy to extract unique patterns in each sequence at different unique levels. A fundamental unique peptide segment with previously defined pattern length, named as primary pattern was extracted at the first step. The rule of thumb for primary pattern lengths is that a shorter length setting for similar protein sequences and a longer length for dissimilar ones. The length of primary pattern in this study is set as 3 residues for hRNaseA protein family. After that Boyer Moore algorithm was performed to efficiently retrieve all primary patterns among all sequences. Each verified fundamental unique peptide segment was analyzed based on its frequencies of appearance, and its representation level of uniqueness was calculated for the merging processes in the next module. The last merging algorithm concatenated these extracted unique peptide segments through a bottom-up approach only if the primary unique peptide segments were overlapped within a sequence. The merged segments were guaranteed with unique features compared to all other protein sequences in the query dataset.

Clustal W2 (http://www.ebi.ac.uk/Tools/msa/clustalw2/) was used to align protein sequences of 13 hRNaseA family members on the basis of automatically progressive alignment mode. Protein sequences were retrieved from UniProtKB (http://www.uniprot.org/). To perform multiple-sequence alignment, gap open and extend penalties were set to 10 and 0.2, respectively. For secondary structure analysis in corresponding HBRs of hRNase1 to hRNase8, tertiary structures of hRNase1, 2, 3, 4, 5, and 7 were collected from protein data bank (PDB, http://www.rcsb.org/pdb/home/home.do), and those of hRNase6 and hRNase8 were simulated by Protein Structure Prediction Server (PS)^2 (http://ps2.life.nctu.edu.tw/) using hRNase7 as a template. In addition, National Center for Biotechnology Information (NCBI) Blast (http://blast.ncbi.nlm.nih.gov/Blast.cgi) was employed to compare sequence correspondent to $HBP_{RNase3(32–41)}$ among nonhuman primate hRNase2s and hRNase3s, as well as human RNaseA superfamily members.

2.2. Cell Line Strains. Beas-2B (ATCC number: CRL-9609) was a human lung bronchial epithelial cell line infected with an adenovirus 12-SV40 virus hybrid (Ad12SV40). Beas-2B cells were cultured in RPMI 1640 medium (Gibco, Invitrogen, USA) supplemented with 56°C heat-inactivated 10% (v/v) fetal bovine serum (FBS) (Gibco, Invitrogen, USA) and 1% (v/v) Glutamine-Penicillin-Streptomycin (Biosera). Wild-type Chinese Hamster Ovary- (CHO-) K1 and mutant cell lines CHO-pgsA-745 (lacking of all GAGs) and CHO-pgsD-677 (lacking of HS) were cultured in Vitacel Ham's F12 K medium (Sigma-Aldrich) supplemented with 10% FBS. Beas-2B and pgs D-677 cell lines were purchased from ATCC. CHO-K1 and CHO-pgsA745 cell lines were kindly provided by Dr. W.-G. Wu and Dr. C.-L. Yang, respectively (Department of Life Science, National Tsing Hua University, Taiwan). The cells were maintained at 37°C in a humidified atmosphere of 5% CO_2.

2.3. Recombinant Protein Purification. Recombinant wild-type hRNase3 with a *C*-terminal His_6 tag was expressed in *Escherichia coli* BL21(DE3) Codon Plus (Novagen), purified by affinity column chromatography, and refolded as previously described [27, 43]. The plasmids containing insert encoding mutant HBR1, HBR2, and HBR3 hRNase3 were generated by QuikChange site-directed mutagenesis with

primer sets mtr1 forward: 5′-TATGCAGCGGCTTGCGCAAACCAAAAT-3′ and mtr1 reverse 5′-TTTGCGCAAGCCGCTGCATAATTGTTA-3′; mtr2 forward 5′-AGGCACGGCGGCGGCGGCGGCATGACAATTGTTGAG-3′ and mtr2 reverse 5′-TGTCATGCCGCCGCCGCCGCCGTGCCTTTACTCCAC-3′; mtr3 forward 5′-ATAGAAGGCGGCGGCGGCGGCGTCTGCATACCTGCA-3′ and mtr3 reverse 5′-GCAGACGCCGCCGCCGCCGCCTTCTATGTAGTTGCA-3′. For each preparation, 10 mL of overnight culture was subcultured into 1 L TB containing 100 μg/mL ampicillin, and grown at 37°C for 6 h. Isopropyl-β-D-thiogalactopyranoside (IPTG) was added to a final concentration of 0.5 mM. Wild-type and mutant hRNase3 were collected from inclusion bodies that were refolded by dialysis in refolding buffer (20 mM Tris, 0.5 M arginine, 0.2 mM GSSG, 2 mM EDTA, 10% glycerol, pH 8.5) at 4°C gently, concentrated by Amicon Ultra-15 (Millipore), and stored in phosphate-buffered saline (PBS).

2.4. Fluorescence-Assisted Carbohydrate Electrophoresis (FACE). Carbohydrates were labeled with 2-aminoacridone (AMAC) according to previous study [44]. The AMAC-labeled carbohydrate and peptide were mixed and incubated at 25°C for 15 min. The complex was then loaded onto 1% agarose gels and electrophoresed in the buffer containing 40 mM Tris-acetic acid, 1 mM EDTA, pH 8.0 for 20 to 30 min. This experiment was performed in dark or under red light to prevent from light exposure. The AMAC labeled probe was observed under UV light (424 nm) and scanned by transilluminator.

2.5. RNase Activity Assay. RNase activity assay of recombinant wild-type and mutant hRNase3 were performed using yeast tRNA (Invitrogen), and bovine RNaseA (USB) was typically RNase and used as positive control. Three hundred microliters of 100 mM sodium phosphate ($NaPO_4$) buffer, pH 7.4, and 500 μL diethylpyrocarbonate- (DEPC-) ddH_2O were mixed including 50 μL of 0.05 μM RNaseA and 5 μM of wild-type and mutant hRNase3, separately. Ten microliters of 5 mg/mL yeast tRNA was added and incubated at 37°C for 0, 5, 10, and 15 min, respectively. Ice-cold 500 μL stop solution (1 : 1 (v/v) 40 mM lanthanum nitrate and 6% perchloric acid) was added and mixed for 10 min to stop reaction. Entire yeast tRNA was suspended by centrifugation at 16,100 ×g at 25°C for 5 min. One hundred microliters of supernatant in each tube was placed on to a 96-well plate. The amount of soluble tRNA in supernatant was determined by UV absorbance at 260 nm.

2.6. MTT Assay. Beas-2B cells (5×10^3 cells/well) were plated in each well of a 96-well plate and allowed to incubate at 37°C for 24 h. Beas-2B cells were incubated with 0, 5, 10, 20, 40, 60, 80, and 100 μM $HBP_{RNase3(32–41)}$ and were incubated at 37°C for 24 h. 3-(4,5-Dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide (MTT) solution (10 μL of 10 mg/mL, Roche) was added to each well, and the cultures were incubated for an additional 4 h. After 24 h, the culture medium was replaced by 100 μL 5 mg/mL MTT for 3 h, then MTT solution was substituted by 100 μL 100% DMSO for 15 min. Finally, the absorbance of the sample was measured at 570 nm.

2.7. Uptake Assay and Western Blotting. Beas-2B cells were seeded into 100 mm dish at 37°C for 24 h and followed by washed with PBS and incubated with 1 μM wild type or mutant hRNase3 in serum-free RPMI 1640 medium at 37°C for 1 h. The cells were washed with PBS and trypsinized for at 37°C 15–20 min to remove surface-associated protein, and cells were lysed with Pro-prep (iNtRon). After electrophoresis, proteins were transferred to PVDF membrane (Pall) using a Bio-Rad Trans-Blot semidry transfer cell. Membranes were blocked with 3% BSA in 0.15 M NaCl, 0.5% (v/v) Tween-20, and 20 mM Tris, pH 7.4 (TBST) followed by primary antibody (anti-His antibody) diluted in 1% BSA/TBST. After washing with TBST, membranes were incubated with a secondary antibody (anti-mouse antibody) diluted in 1% BSA/TBST. Bound antibody was detected using the enhanced chemiluminescence detection system (Pierce).

2.8. Synthetic Peptides. $HBP_{RNase3(32–41)}$ (NYRWRCKNQNK), $HBP_{RNase3(71–80)}$ (NNCHRSRFRV) and $HBP_{RNase3(96–105)}$ (TYADRPGRRF) were synthesized by Genemed Synthesis Inc. (USA). All peptides were purified by analytical high-pressure liquid chromatography to a purity exceeding 90%. The identity of peptides was confirmed by matrix-assisted laser desorption ionization time of flight mass spectrometry.

2.9. Cell-Based Enzyme-Linked Immunosorbent Assay (cELISA). Cells (2×10^4 cells/well) were seeded in a 96-well black plate and incubated with 5% CO_2 at 37°C for 24 h. The fluorescein isothiocyanate- (FITC-) labeled $HBP_{RNase3(32–41)}$, $HBP_{RNase3(71–80)}$, and $HBP_{RNase3(97–106)}$ were diluted to 0, 1, 5, and 10 μM in medium for 1 h. After wash with 100 μL PBS, the plate was fixed by 2% (w/v) paraformaldehyde (PFA) in PBS for 15 min, and 100 μL 50 mM ammonium chloride in PBS was added to quench fluorescence. The plate was washed with 100 μL PBS, and 2% (w/v) BSA in PBS was added to block at room temperature for 1.5 h. The signals were measured using 488 nm laser with standard 530 nm ± 30 nm bandpass emission filter (Omega Optical, Brattleboro, VT, USA). In each set of the assays, analyses were carried out in duplicate or triplicate.

2.10. Lipid Overlay Blots. SphingoStrips (Invitrongen catalog no. S23753) and PIP Strips (Invitrongen catalog no. P23751) membranes were blocked at 37°C for 1 h with Tris-buffered Saline/0.05% Tween-20 (TBST) containing 3% (w/v) fatty acid free BSA. Membranes were incubated with 1 μg/mL hRNase3 or 0.5 μg/mL FITC-$HBP_{RNase3(32–41)}$ at 37°C for 2 h separately, the membranes were washed once using 0.1% BSA/TBST with a gentle shaking for 10 min. hRNase3 binding to lipids was probed by primary antibody (mouse anti-6His 1 : 5000) for 1 h. After wash, anti-mouse antibody conjugated with horseradish peroxidase (HRP) was used as secondary antibody (1 : 5000) in 1% BSA/TTBS. Finally, the

hRNase3
RPPQFTRAQWFAIQHISLNPPRCTIAMRAINNYRWRCKNQNTFLRTTFANVVNVCGN
QSIRCPHNRTLNNCHRSRFRVPLLHCDLINPGAQNISNCTYADRPGRRFYVVACDNR
DPRDSPRYPVVPVHLDTTI

Figure 1: Unique peptide motifs in hRNase3 revealed from 13 hRNaseA members by Reinforced Merging for Unique Segments (ReMUS) system. Blue and light blue segments represent unique peptide motifs of hRNase3 compared to other 12 hRNaseA members. Three unique motifs ^{34}RWRCK^{38}, ^{75}RSRFR^{79}, and ^{101}RPGRR^{105}, representing, respectively, HBR1, HBR2, and HBR3 in hRNase3 are highlighted in yellow.

immunoreactive bands were visualized by enhanced chemiluminescence (USA). For FITC-HBP$_{\text{RNase3(32–41)}}$ analysis, the membrane strips were initially incubated with FITC-HBP$_{\text{RNase3(32–41)}}$ followed by PBS wash and finally imaged using 488-nm laser with standard 530 nm ± 30 nm bandpass emission filter.

3. Result

3.1. In Silico Analysis of Unique Peptide Regions in hRNaseA Superfamily. To predict unique peptide regions possibly involved in multifunctions of hRNase3, ReMUS system was employed to analyze sequences of hRNase3 and the other 12 members of hRNaseA family. Eleven unique peptide motifs including HISLNPPR, RCTIAMRA, NYRWRC, SIRCPHNRTLNNC, RSRFRVP, PLLHCD, DLINP, PGAQN, NCTYADRPGRRFYV, DPRDSPRY, and LDTTI in hRNase3 were identified as shown in blue and light blue in Figure 1. Among which three unique segments rich in positively charged amino acids were denoted as putative HBRs including ^{34}RWRCK^{38} (HBR1), ^{73}RSRFR^{77} (HBR2), and ^{101}RPGRRR^{105} (HBR3). Since heparin binding activity of HBR1 has been previously reported [27], the presence of 3 HBRs might possibly correlate with stronger heparin binding features of hRNase3 than other hRNaseA family members. Subsequently, Clustal W2 was applied to compare primary sequence of hRNase3 with the other hRNaseAs and alignment of putative HBRs of hRNase3 with correspondent segments of the other 12 hRNaseAs. Figure 2 revealed that HBR1 in hRNase3 was 60% identical to the corresponding segments of hRNase1, hRNase2, hRNase7, and hRNase8, but these HBRs were not conserved with any of the other hRNase family members, suggesting that these three HBRs might account for unique functions of hRNase3. Therefore, seven mutant hRNase3 constructs were generated by site-directed mutagenesis with selective alanine replacement in each HBR in order to investigate unique functions of HBR1, HBR2 and HBR3 in hRNase3 (Supplementary Figure 1 available online at http://dx.doi.org/10.1155/2013/170398).

3.2. RNase Activity of Wild-Type and Mutant RNase3. After 0.05 μM bovine RNaseA, 5 μM wild-type or mutant hRNase3 was incubated with 50 μg tRNA at 37°C for 0, 5, 10, and 15 min, the amount of digested ribonucleotides was examined by UV absorbance at 260 nm. Figure 3(a) showed that RNase activity of HBR1-mt RNase3 was significantly reduced and that of HBR3-mt RNase3 was 15% less than wild type hRNase3. Besides, the RNase activity of HBR2-mt RNase3 was comparable to that of wild-type hRNase3. Moreover, double and triple mutation abolished hRNase activities of HBR12-, HBR13- and HBR123-mtRNase3 expect HBR23-mtRNase3 (Figure 3(b)). These results suggested that HBR1 of hRNase3 played a critical role in ribonucleolytic activity, mainly due to the presence of a catalytic residue Lys in the sequence.

3.3. Endocytosis Activity of Wild-Type and Mutant hRNase3 to Beas-2B Cells. To determine the influence of different HBRs on endocytosis activity of hRNase3, intracellular uptake assay was performed. Beas-2B cells were treated with wild type or mutant hRNase3 at 37°C for 1 h, followed by trypsin digestion for 15 min to remove surface-bound recombinant proteins before being analyzed by western blotting. When 40 μg of total cell lysates were examined with an exposure time of 1 min, only HBR2-mtRNase3 and HBR3-mtRNase3 could be detected in cytosol of Beas-2B cells (Figure 4, lanes 3 and 4). In addition, none of the double and triple HBR mutants of hRNase3 was able to enter Beas-2B cells (Figure 4, lanes 5, 6, 7, and 8). These results indicated that the importance of HBRs associated with endocytosisactivity of hRNase3 to Beas-2B cells in increasing order was HBR3, HBR2, and HBR1.

3.4. Cytotoxicity of Wild-Type and Mutant hRNase3 to Beas-2B Cells. Beas-2B cells were incubated individually with 15 μM of wild-type or mutant hRNase3 in serum-free medium at 37°C for 48 h followed by MTT assay. The cell viability of PBS treatment to Beas-2B cells was set as 100% to normalize that of different protein treatment. Figure 5 revealed that the viability of wild-type hRNase3-treated cells decreased to 50%, while that of HBR1-, HBR2-, HBR3-, HBR12-, HBR13-, HBR23-, and HBR123-mtRNase3-treated cells increased to 69%, 60%, 101%, 73%, 96%, 88%, and 97%, respectively. The cytotoxicity of HBR3-mtRNase3, HBR13-mtRNase3, HBR23-mtRNase3, and HBR123-mtRNase3 apparently diminished as compared to that of wild-type hRNase3, suggesting that HBR3 mutation played a major role in loss of cytotoxicity of hRNase3. It should be noted that HBR3 was located on β sheet 6 of hRNase3, hence mutation of HBR3 to alanine stretch might possibly lead to conformational change and subsequent of functional variation.

3.5. Binding Activity of Wild Type and Mutant hRNase3 to LMWH. The influence of wild type and mutant hRNase3 on LMWH binding activity was illustrated in Figure 6. Initially, AMAC-labeled LMWH was coincubated with wild-type or mutant hRNase3 individually at a molar ratio of protein to LMWH of 0.3. Free probe was separated by 1% gel electrophoresis. Relative intensities of free probes of wild type, HBR1-, HBR2-, HBR3-, HBR12-, HBR13-, HBR23-, and HBR123-mt RNase3 apparently decreased to, respectively, 17%, 31%, 15%, 9%, 76%, 59%, 76%, and 81%. The binding activity of HBR1-mtRNase3 to LMWH decreased 14% as compared to that of wild type hRNase3 (Figure 6, lanes 2

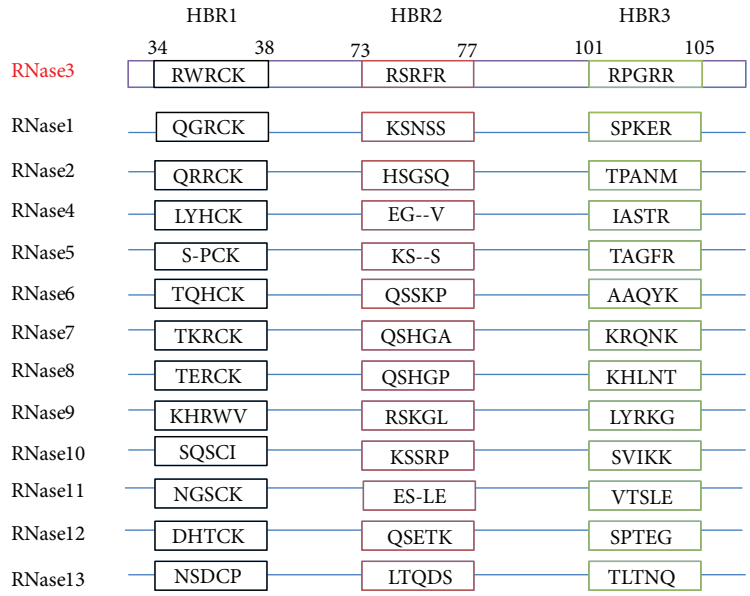


Figure 2: Sequences of HBR1, HBR2, and HBR3 in hRNase3 and corresponding regions of other 12 hRNaseA members. Sequences of hRNase3 and other hRNaseA superfamily members are aligned using ClustalW2 software. Putative HBR1, HBR2, and HBR3 separately located on residues 34–38, 73–77, and 101–105 are predicted from hRNase3. Residues in black, red, and green boxes indicate corresponding sequence motifs aligned with, respectively, HBR1, HBR2, and HBR3 in all hRNaseA superfamily members.

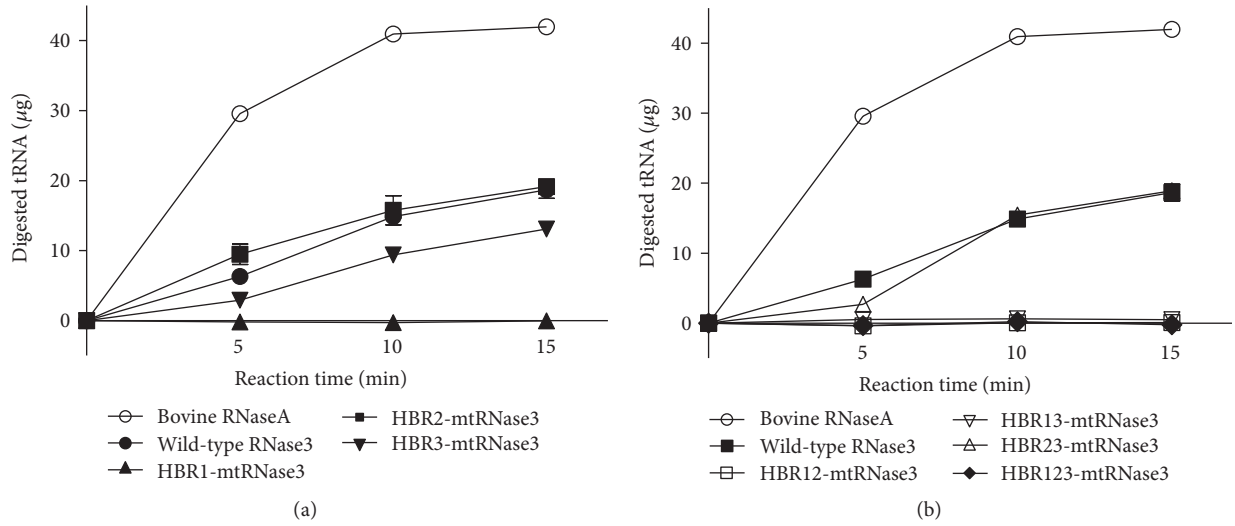


Figure 3: RNase activity of wild-type and mutant hRNase3. Five mg of yeast tRNA was added to 0.05 μM bovine RNaseA and 5 μM of (a) HBR1-mtRNase3, HBR2-mtRNase3, and HBR3-mtRNase3 and (b) HBR12-mtRNase, HBR13-mtRNase3, HBR23-mtRNase3, and HBR123-mtRNase3 separately to examine the RNase activity. The amount of digested ribonucleotides in supernatant was detected by monitoring OD_{260} and RNase activity of bovine RNaseA was set as a positive control.

and 3), while HBR2- and HBR3-mt RNase3 did not show much difference in LMWH binding activity (Figure 6, lanes 4 and 5). However, the binding activity of double- and triple-mutant hRNase3 to LMWH decreased more significantly, especially when HBR1 was mutated (Figure 6, lanes 6, 7, 8, and 9). Therefore, in terms of facilitating hRNase3 binding to LMWH, the importance of three HBRs in decreasing order was HBR1, HBR2, and HBR3.

3.6. Binding Activity of Synthetic Peptides FITC-$HBP_{RNase3(32–41)}$, FITC-$HBP_{RNase3(71–80)}$, and $HBP_{RNase3(97–106)}$ to Beas-2B Cells. FITC-labeled $HBP_{RNase3(32–41)}$ (NYRWRCKNQNK), $HBP_{RNase3(71–80)}$ (NNCHRSRFRV) and $HBP_{RNase3(97–106)}$ (TYADRPGRRF), peptides containing, respectively, HBR1 (^{34}RWRCK^{38}), HBR2 (^{73}RSRFR^{77}) and HBR3 (^{101}RPGRR^{105}) sequences were synthesized to investigate the cellular binding activity

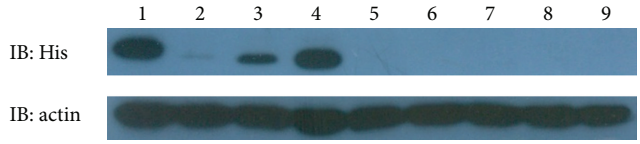


Figure 4: Endocytosis of wild-type and mutant hRNase3 to Beas-2B cells. Endocytosis activity of hRNase3 and its HBR mutants were assayed by western blotting employing 1 : 5000 dilution of anti-His-antibody. Actin was used to normalize the hRNase3-6His immunoblot signal. Forty micrograms of total protein in cell lysates were separated by 15% SDS-PAGE. IB: immunoblot; lane 1: wild-type hRNase3; lane 2-HBR1-mtRNase3; lane 3-HBR2-mtRNase3; lane 4-HBR3-mtRNase3; lane 5-HBR12-mtRNase3; lane 6-HBR13-mtRNase3; lane 7-HBR23-mtRNase3; lane 8-HBR123-mt RNase3; lane 9-cell only.

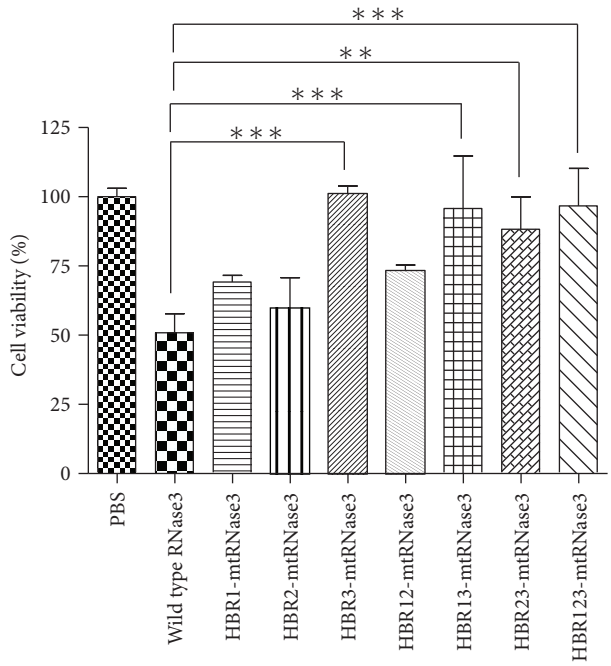


Figure 5: Cytotoxicity of wild-type and mutant hRNase3 to Beas-2B cells. Cytotoxicity of hRNase3 and its HBR mutants to Beas-2B cells were assessed by MTT assay. Cell viability of Beas-2B cells with PBS treatment was set as 100%. Results are presented as mean ± SD ($n = 3$). $^{**}P < 0.01$, $^{***}P < 0.005$.

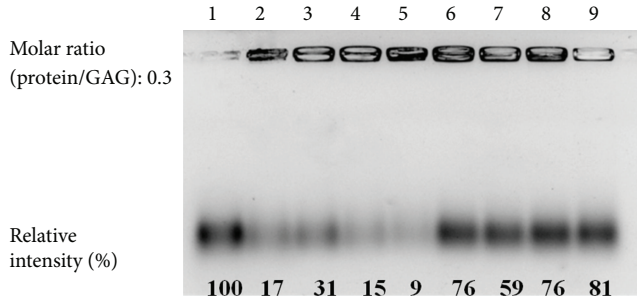


Figure 6: Binding activity of wild-type and mutant hRNase3 to LMWH. Heparin binding activity of hRNase3 and its HBR mutants were assessed by FACE. The value of labeled LMWH in the absence of protein was measured and set as 100%. Relative intensity of LMWH signal of each protein was normalized to LMWH only signal and was marked at the bottom; lane 1-LMWH only; Lane 2-hRNase3; lane 3-HBR1-mtRNase3; lane 4-HBR2-mtRNase3; lane 5-HBR3-mtRNase3; lane 6-HBR12-mtRNase3; lane 7-HBR13-mtRNase3; lane 8-HBR23-mtRNase3; lane 9-HBR123-mt RNase3.

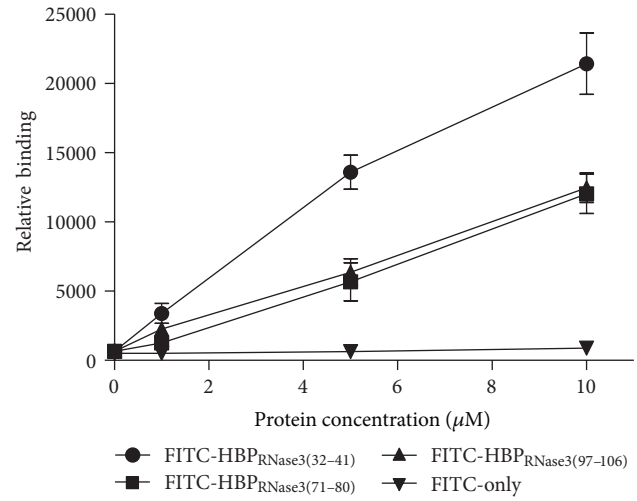


Figure 7: Binding activity of FITC-HBP$_{RNase3}$ peptides to Beas-2B cells. Cellular binding activity of HBPs derived from hRNase3 including FITC-HBP$_{RNase3(32–41)}$, FITC-HBP$_{RNase3(71–80)}$, and FITC-HBP$_{RNase3(97–106)}$ were assayed using cELISA. The level of bound FITC signal was set as negative control. Data represented the means of triplicate incubations. Error bars showed standard deviations (SD) among triplicate experiments.

by cELISA. Figure 7 indicated that the binding activity of FITC-HBP$_{RNase3(32–41)}$, FITC-HBP$_{RNase3(71–80)}$, and HBP$_{RNase3(97–106)}$ to Beas-2B cells increased with elevated peptide concentration ranging from 1 μM to 10 μM in a dose-dependent manner, the binding activity of FITC-HBP$_{RNase3(32–41)}$ was 2 times stronger than that of FITC-HBP$_{RNase3(71–80)}$ and HBP$_{RNase3(97–106)}$. Since HBR1 was a stronger heparin binding site than HBR2 and HBR3, its multiple functions were further investigated.

3.7. Cellular Binding Activity of hRNase3 and HBP$_{RNase3(32–41)}$. A series of wild-type and mutant CHO cell lines with specific defect in HS or GAG biosynthesis were employed to study the binding target ofhRNase3 and HBP$_{RNase3(32–41)}$ on plasma membrane. Figure 8(a) showed that 1000 nM hRNase3 binding to CHO-pgsD677 and CHO-pgsA745 cells was 50% lower than that of wild-type CHO-K1 cells. Similarly, 5 μM FITC-HBP$_{RNase3(32–41)}$ binding to CHO-pgsD677 and CHO-pgsA745 cells significantly decreased, respectively, 40% and 50% as compared to that of wild-type CHO-K1cells (Figure 8(b)), and FITC-only was used a negative control and its relative binding intensities to each cell line as only 10%. While lacking of HS or GAGs on cell surface, relative hRNase3 and HBP$_{RNase3(32–41)}$ binding intensities to both CHO-pgsD677 and CHO-pgsA745 cells dropped about 50%. These results indicated that the presence

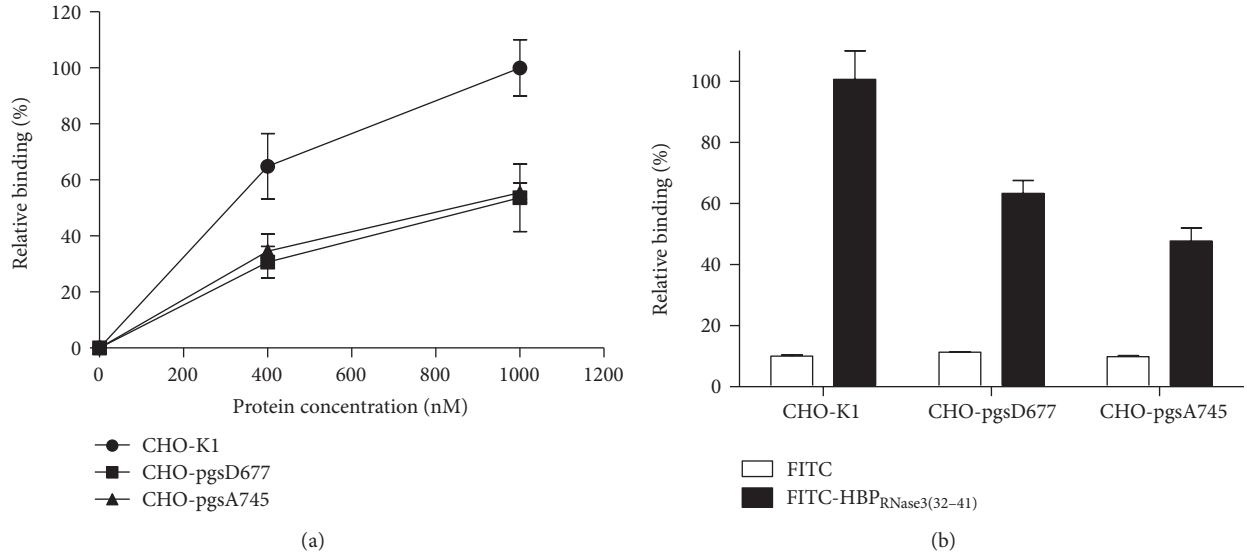


FIGURE 8: Binding activity of hRNase3 and FITC-HBP$_{RNase3(32–41)}$ to wild-type and mutant CHO cell lines. Cellular binding activity of hRNase3 (a) and FITC-HBP$_{RNase3(32–41)}$ (b) to CHO-K1, CHO-pgsD677, and CHO-pgsA745 cells were assayed using cELISA. The amount of 1000 nM hRNase3 and 5 μM FITC-HBP$_{RNase3(32–41)}$ bound to wild-type CHO-K1cell was set as 100% binding, respectively. The data represented the mean value of triple independent experiments and the error bar was shown as standard deviation (SD).

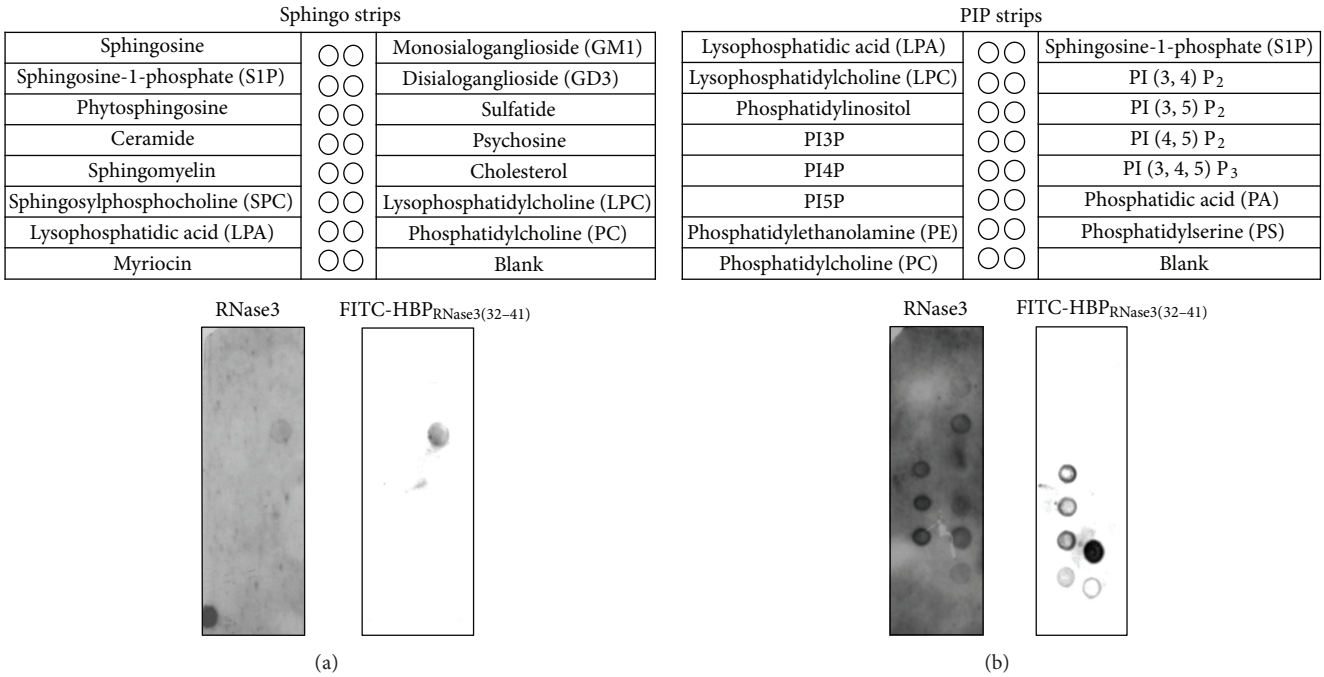


FIGURE 9: Interaction of hRNase3 and FITC-HBP$_{RNase3(32–41)}$ with membrane lipid fractions. Interaction of hRNase3 and FITC-HBP$_{RNase3(32–41)}$ on lipid overlay blot of SpingoStrips membranes (a) and PIP Strips membranes (b). Membranes were incubated with 1 μg/mL hRNase3 or 0.5 μg/mL FITC-HBP$_{RNase3(32–41)}$ at 37°C for 2 h, separately. The immunoreactive blot of hRNase3 treatment was visualized by enhanced chemiluminescence, and the FITC-HBP$_{RNase3(32–41)}$ treatment was using 488 nm laser with standard 530 nm ± 30 nm bandpass emission filter.

of HS on the cell surface was most essential for molecular interaction with hRNase3 and HBP$_{RNase3(32–41)}$. In addition, cell surface components, possibly lipid moiety on A745 cells, might be involved in residual hRNase3 and HBP$_{RNase3(32–41)}$ binding activities.

3.8. Identification of Specific Lipids Interacting with RNase3 and HBP$_{RNase3(32–41)}$. hRNase3 has been associated with lipid membrane destabilization [45]. To investigate whether any other membrane-associated negatively charged moieties other than HS was involved in hRNase3 and HBP$_{RNase3(32–41)}$

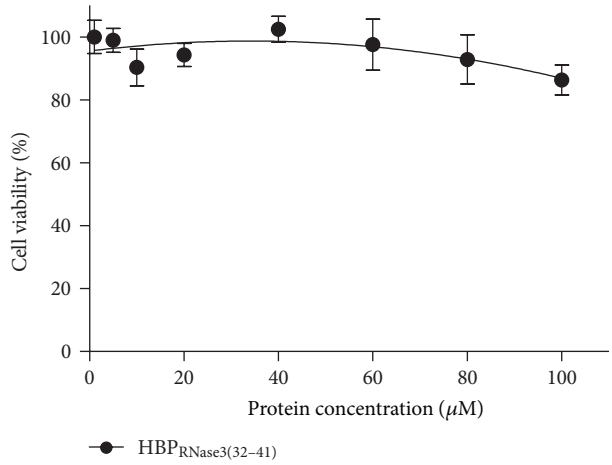


FIGURE 10: Cell viability of Beas-2B cells upon treatment with HBP$_{RNase3(32–41)}$. Cytotoxicity of HBP$_{RNase3(32–41)}$ to Beas-2B cell was assessed by MTT assay. Beas-2B cells were treated with various concentrations of HBP$_{RNase3(32–41)}$. The amount of cell viability of PBS treated Beas-2B cells was set as 100%.

binding, overlay blots (SphingoStrips and PIPStrips) containing 100 pmoles of 26 different biologically active sphingolipids and glycerophospholipids were examined [46]. Using SphingoStrips membrane, a specific interaction between sulfatide, a sulfated glycosphingolipid expressed on the cell surface [47], and refold hRNase3 (Figure 9(a), left panel) as well as FITC-HBP$_{RNase3(32–41)}$ (Figure 9(a), right panel) was observed. For the PIP Strips membrane, hRNase3 interacted with PI3P, PI4P, PI5P, PI (3,4) P_2, PI (3,5) P_2, PI (3,4,5) P_3, PA, and PS (Figure 9(b), left panel). FITC-HBP$_{RNase3(32–41)}$ interacted with PI3P, PI4P, PI5P, PE, PA, and PS (Figure 9(b), right panel). These results indicated that FITC-HBP$_{RNase3(32–41)}$ possessed quite similar lipid binding pattern to hRNase3, and the HBP$_{RNase3(32–41)}$ involved in hRNase3 binding to plasma membrane not only through HS binding but also through direct molecular interaction with phospholipids. Our data suggested that hRNase3 interacted with more lipids than HBP$_{RNase3(32–41)}$, presumably due to more diverse lipid binding regions on hRNase3. Besides, variation in the exposed level and conformation of ^{32}NYRWRCKNQN^{41} sequence between refold hRNase3 and HBP$_{RNase3(32–41)}$ might also account for the difference in lipid recognition.

3.9. Cytotoxicity of hRNase3 and HBP$_{RNase3(32–41)}$ to Beas-2B Cells. Cytotoxicity of hRNase3 towards Beas-2B cells has been reported in previous study [27]. To examine the cytotoxic effect of HBP$_{RNase3(32–41)}$, Beas-2B cells were incubated with 0, 5, 10, 20, 40, 60, 80, and 100 μM HBP$_{RNase3(32–41)}$ in serum-free medium at 37°C for 24 h and cell viability was measured by MTT assay. Figure 10 revealed that no cytotoxicity of HBP$_{RNase3(32–41)}$ was detected, even if high concentration of 100 μM was applied. Thus, although HBP$_{RNase3(32–41)}$ conserved key heparin binding activity of hRNase3, it possessed no cytotoxic effect to mammal cell.

4. Discussion

hRNaseA superfamily members share diverse protein identities to hRNase3 even though they contain conserved 3-dimensional structures and enzymatic functions. In addition to RNase activity, a variety of biological features including immune-regulatory, cytotoxic, antimicrobial, antitumor, and heparin/HS binding activities have been reported. In terms of RNase activities, hRNases1-to-8 have obvious catalytic activities; however, such function is unidentified in hRNases9-to-13 and remains to be elucidated. As for cytotoxicity, hRNases1-to-5 are harmful to various cells or organisms, but hRNases6-to-13 are indistinct. In terms of lipid binding activity, hRNase2, hRNase3, and hRNase7 have been reported to interact with lipids [30–32], while that of the other hRNaseA family members are not well studied. As for antimicrobial activity, hRNase2, hRNase3, hRNase5, hRNase7, and hRNase8 are harmful to microorganisms (Table 1). In addition, hRNase2 and hRNase3 have been elucidated to mediate immune responses [31, 32]. Taken together, comparison of hRNaseA family members indicates that only hRNase2 has similar sequences and functions to hRNase3 (Table 1).

Structural analysis reveals that all hRNaseA family members share very similar secondary structures in three putative HBRs (Table 2). Since hRNase7 shares high primary sequence identity to hRNase6 and hRNase8, 58% and 75%, respectively, its structure was used as a template for structure simulation of hRNase6 and hRNase8 which have no resolved 3D structures yet. The HBR1 and HBR3 in all hRNaseA family members are present, respectively, in loop and β-strand conformation. The secondary structure for HBR2 of hRNase4 is β-strand, and that of the others is loop conformation. Hence, these results suggested that sequence composition, rather than secondary structure contents of each HBR-like segment in hRNaseA family members, is crucial for differential heparin binding activities. Among all hRNaseA superfamily members, hRNase2, hRNase3, and hRNase7 possess conventional heparin binding motifs in several HBRs. However, only hRNase2, hRNase3, and hRNase5 have been reported to demonstrate heparin binding activities [48, 49]. Here, three unique functional peptides encoded HBRs have been predicted in hRNase3 by ReMUS and demonstrated heparin binding properties at both molecular and cellular levels, and the correspondent HBR1 of hRNase2 has been identified with heparin binding features too. Interestingly, in the primary sequence hRNase5 positively charged residue-rich regions, that is, ^{31}RRR^{33} and R^{70} have been reported to involve heparin binding by site-directed mutagenesis [49], and they are located pretty close to putative HBR1 and HBR2 of hRNase5 in this study. Finally, hRNase7 has been reported to be purified through a heparin affinity column despite unclear heparin binding mechanism, indicating that hRNase7 also possesses heparin binding potency. This finding will contribute to further understanding of protein-ligand interaction in hRNaseA members.

hRNase3 is a ribonuclease and its antimicrobial function has been shown to be dependent on its enzymatic activity [5, 19, 25]. Our study has identified three HBRs including

Table 1: Specific functions of hRNaseA superfamily.

Function	RNase							
	1	2	3	4	5	6	7	8
Protein identity	30%	67%	100%	28%	32%	43%	39%	39%
RNase activity	O	O	O	O	O	O	O	O
Cytotoxicity	O	O	O	O	O	ND	ND	ND
Heparin binding motif	X	O	O	X	O	O	O	X
Lipid binding activity	ND	O	O	ND	ND	ND	O	ND
Antimicrobial activity	ND	O	O	ND	O	ND	O	O
Inflammatory mediators	ND	O	O	ND	ND	ND	ND	ND
Reference	[29]	[30]	[31, 32]	[33, 34]	[35, 36]	[37]	[38, 39]	[40, 41]

Note: O, X, and ND, respectively, represent active, inactive, and not determined.

Table 2: Secondary structures of correspondent HBRs in hRNaseA members.

	HBR1	HBR2	HBR3	PDB number
hRNase1	Loop	Loop	β-strand	2K11
hRNase2	Loop	Loop	β-strand	2C01
hRNase3	Loop	Loop	β-strand	2KB5
hRNase4	Loop	β-strand	β-strand	1RNF
hRNase5	Loop	Loop	β-strand	1H53
hRNase6	Loop	Loop	β-strand	2HKY*
hRNase7	Loop	Loop	β-strand	2HKY
hRNase8	Loop	Loop	β-strand	2HKY*

The structures of hRNase1, 2, 3, 4, 5, and 7 are collected from protein data bank (PDB), and those of RNase6, 8 were simulated by database $(PS)^2$—Protein Structure Prediction Server (http://ps2.life.nctu.edu.tw/).
*Denotes simulated structures using hRNase7 as template.

^{34}RWRCK^{38}, ^{73}RSRFR^{77}, and ^{101}RPGRRR^{105}, respectively, located on loop3, loop5, and strand β4 of hRNase3 to interact with heparin. Key roles including heparin binding, cytotoxic, endocytic, and lipid binding activities were contributed by these major HBRs and have been demonstrated. hRNase3 can also modulate Beas-2B cells to release TNF-α leading to apoptosis facilitated by first step attachment on cell surface GAGs especially HS [26]. In 2010, Garcia-Mayoral group reported that the major HS binding site on hRNase3 was located on a cavity composed by A^8-Q^{14} in helix α1, Y^{33}-R^{36} in loop3, Q^{40}-L^{44} in strand β1, and H^{128}-D^{130} in strand β6 [18]. Here HBR1 mutants significantly diminished their RNase activity due to replacement of the crucial catalytic residue Lys^{38} to Ala, in addition, it also decreased heparin binding and subsequent endocytosis by replacing three basic residues, Arg^{34}, Arg^{36}, and Lys^{38} and aromatic residue Trp^{35} responsible for binding and disrupting microbial membrane [50]. Mutation in HBR3 (^{101}RPGRR^{105}) significantly decreased the cytotoxicity of hRNase3 to Beas-2B cells, revealing that the three cationic residues Arg^{101}, Arg^{104}, and Arg^{105}in HBR3 might play a crucial role in exterminate Beas-2B cells.

$HBP_{RNase3(31–41)}$ segment was demonstrated to be involved in multiple functions of hRNase3 (Table 3). Alignment with primate RNase3 sequences has revealed that the sequence of $HBP_{RNase3(32–41)}$ is conserved only in RNaseA family members of higher primates such as *G. Gorilla* and *P. troglodytes*, the closest living relatives of human. Interestingly, higher primates and the closest living relatives of humans, *P. troglodytes* and *G. Gorilla* [51] showed 100% sequence identity in corresponding regions to the $HBP_{RNase3(32–41)}$ motif of *H. sapiens* RNase3; while *M. fascicularis* and *M. nemestrina* showed 80% sequence identity with $HBP_{RNase3(32–41)}$ motif of *H. sapiens* RNase3 but 100% identity with *H. sapiens* RNase2. In addition, the motif sequence of *P. pygmaeus* was 70% and 90% identical to that of hRNase3 and hRNase2, respectively (Table 4). Our results strongly supported the notion that hRNase3 was generated from an RNase2/RNase3 precursor gene about 30 million years ago, at an evolutionary rate among the highest primate genes [52].

To date, 13 members of the RNaseA superfamily have been identified in humans; however, the functions of newly identified human hRNases9–13 remain unclear [53]. Blast *analysis* of $HBP_{RNase3(32–41)}$ motif among hRNase3 and other hRNaseA members was shown in Table 5, in which only hRNase2 and hRNase8 showed, respectively, 80% and 50% sequence identity, while the others showed lower than 50% identity with the $HBP_{RNase3(32–41)}$ of hRNase3. In summary, HBP_{RNase3} motif was not a conserved motif among hRNaseA superfamily, but a specific motif being present only in higher primates.

Herein we reported that $HBP_{RNase3(32–41)}$ accounted for major cellular binding activity than $HBP_{RNase3(71–80)}$ and $HBP_{RNase3(97–106)}$. Interestingly, although binding of $HBP_{RNase3(32–41)}$ was severely impaired in CHO-pgsD677 cells which had no HS but expressed 3-fold more CS than wild-type CHO-K1 cells, residual cellular binding activities of $HBP_{RNase3(32–41)}$ to CHO-pgsD677 and CHO-pgsA745 cells strongly implied that $HBP_{RNase3(32–41)}$ possessed certain interaction to cell surface CS. In addition, both hRNase3 and $HBP_{RNase3(32–41)}$ showed quite similar binding activities to membrane lipids.

5. Conclusion

In this study, we identify three functionally important HBRs in hRNase3, including ^{34}RWRCK^{38} (HBR1), ^{75}RSRFR^{79} (HBR2), and ^{101}RPGRR^{105} (HBR3). HBR1 (^{34}RWRCK^{38})

Table 3: Correlation of HBR motifs and characteristic functions of hRNase3.

Location	Function					
	RNase	Cytotoxicity	Cell binding activity	HS binding activity	Endocytic activity	Lipid binding activity
RNase3	O	O	O	O	O	O
HBR1-mt RNase3	No activity	Decrease 50%	Increase 50%	Decrease 40%	No activity	
HBR2-mt RNase3	Similar	Similar	Decrease 20%	Similar	Decrease 50%	
HBR3-mt RNase3	Decrease 30%	No activity	Decrease 20%	Similar	Similar	
$HBP_{RNase3(32–41)}$	X	X	O	O	ND	O

Note: O, X, and ND, respectively, represent active, inactive, and not determined.

Table 4: Comparison of $HBP_{RNase3(32–41)}$ motif among hRNase3 and other primate RNase3s and RNase2s.

Protein	Organism	Protein identity	$HBP_{RNase3(32–41)}$ identity	Sequence
RNase3	*Homo sapiens*	100%	100%	NYRWRCKNQN
	Pan troglodytes	97%	100%	NYRWRCKNQN
	Gorilla gorilla	97%	100%	NYRWRCKNQN
	Macaca fascicularis	88%	80%	NYQRRCKNQN
	Macaca nemestrina	88%	80%	NYQRRCKNQN
	Pongo pygmaeus	88%	70%	NYQRRCKDQN
RNase2	*Homo sapiens*	67%	80%	NYQRQCKNQN
	Pan troglodytes	67%	80%	NYQRQCKNQN
	Gorilla gorilla	69%	80%	NYQRQCKNQN
	Macaca fascicularis	67%	80%	NYQRQCKNQN
	Macaca nemestrina	66%	80%	NYQRQCKNQN
	Pongo pygmaeus	68%	70%	NFQRRCKNQN

Sequence identity was calculated employing National Center for Biotechnology Information Blast (NCBI Blast: http://blast.ncbi.nlm.nih.gov/Blast.cgi).

Table 5: Comparison of $HBP_{RNase3(32–41)}$ motifs among human RNaseA superfamily members.

Name	Accession number	Protein identity	$HBP_{RNase3(32–41)}$ identity	Sequence
RNase1	*P07998*	30%	40%	MTQGRCKPVN
RNase2	*P10153*	67%	80%	NYQRRCKNQN
RNase3	*P12724*	100%	100%	NYRWRCKNQN
RNase4	*P34096*	28%	30%	MTLYHCKRFN
RNase5	*P03950*	32%	30%	LTSP-CKDIN
RNase6	*Q93061*	43%	20%	KYFGRSLELY
RNase7	*Q9H1E1*	39%	40%	KHTKRCKDLN
RNase8	*Q8TDE3*	39%	50%	KYTERCKDLN
RNase9	*P60153*	23%	20%	YYKHRWVAEH
RNase10	*Q5GAN6*	26%	10%	EPSQSCIAQY
RNase11	*Q5GAN5*	30%	30%	EANGSCKWSN
RNase12	*Q5GAN4*	23%	20%	EPDHTCKKEH
RNase13	*Q5GAN3*	25%	10%	MQNSDCPKIH

Sequence identity was calculated employing National Center for Biotechnology Information Blast (NCBI Blast: http://blast.ncbi.nlm.nih.gov/Blast.cgi).

in hRNase3 is required for RNase activity and serves as a major heparin binding site. HBR2 (^{73}RSRFR^{77}) contributes to both cell binding and endocytic activities, and HBR3 (^{101}RPGRR^{105}) plays an important role in cytotoxicity. Moreover, a noncytotoxic HBR1-derived peptide prefers to interact with negatively charged molecules including heparan sulfate, chondroitin sulfate and lipids present on cell surface. Understanding of the roles of key functional residues of hRNase3 in ribonucleolytic, heparin binding, cellular binding, endocytic, cytotoxic, and lipid binding activities provides informative correlation among sequence, structure, and functional features of hRNaseA family members. This

finding will contribute to further investigation of molecular mechanisms and multiple functions of hRNaseA family in general.

Authors' Contribution

P.-C. Lien and P.-H. Kuo contributed equally to this work.

Acknowledgments

The authors thank Mr. Ta-Jen Hung, Dr. Chao-Sheng Cheng, and Dr. Wen-Chi Cheng for critical comments and proofreading. No other potential conflict of interests relevant to this paper was reported. This work was supported by National Tsing Hua University (NTHU100N7051E1 and NTHU101N2051E1), National Science Counsil (NSC101-2622-B-007-001-CC1 and NSC101-2325-B-007-002), and Chang-Gung Memorial Hospital-National Tsing Hua University Joint Research Program (NTHU100N2710E1 and NTHU100N2711E1) to M. D.-T. Chang, and National Science Council (NSC98-2320-B-007-003-MY3) to Y.-K. Lai. P.-H. Kuo and C.-J. Chen were awarded a scholarship sponsored by Shen's Culture and Education Foundation.

References

[1] J. J. Beintema and R. G. Kleineidam, "The ribonuclease A superfamily: general discussion," *Cellular and Molecular Life Sciences*, vol. 54, no. 8, pp. 825–832, 1998.

[2] K. D. Dyer and H. F. Rosenberg, "The RNase a superfamily: generation of diversity and innate host defense," *Molecular Diversity*, vol. 10, no. 4, pp. 585–597, 2006.

[3] S. Cho, J. J. Beintema, and J. Zhang, "The ribonuclease A superfamily of mammals and birds: identifying new members and tracing evolutionary histories," *Genomics*, vol. 85, no. 2, pp. 208–220, 2005.

[4] H. F. Rosenberg, "The eosinophil ribonucleases," *Cellular and Molecular Life Sciences*, vol. 54, no. 8, pp. 795–803, 1998.

[5] M. T. Rugeles, C. M. Trubey, V. I. Bedoya et al., "Ribonuclease is partly responsible for the HIV-1 inhibitory effect activated by HLA alloantigen recognition," *AIDS*, vol. 17, no. 4, pp. 481–486, 2003.

[6] S. Sorrentino, "The eight human "canonical" ribonucleases: molecular diversity, catalytic properties, and special biological actions of the enzyme proteins," *FEBS Letters*, vol. 584, no. 11, pp. 2194–2200, 2010.

[7] T. Maeda, M. Kitazoe, H. Tada et al., "Growth inhibition of mammalian cells by eosinophil cationic protein," *European Journal of Biochemistry*, vol. 269, no. 1, pp. 307–316, 2002.

[8] J. E. Lee and R. T. Raines, "Cytotoxicity of bovine seminal ribonuclease: monomer versus dimer," *Biochemistry*, vol. 44, no. 48, pp. 15760–15767, 2005.

[9] K. J. Hamann, G. J. Gleich, J. L. Checkel, D. A. Loegering, J. W. McCall, and R. L. Barker, "*In vitro* killing of microfilariae of *Brugia pahangi* and *Brugia malayi* by eosinophil granule proteins," *Journal of Immunology*, vol. 144, no. 8, pp. 3166–3173, 1990.

[10] K. J. Hamann, R. L. Barker, D. A. Loegering, and G. J. Gleich, "Comparative toxicity of purified human eosinophil granule proteins for newborn larvae of *Trichinella spiralis*," *Journal of Parasitology*, vol. 73, no. 3, pp. 523–529, 1987.

[11] L. V. Hooper, T. S. Stappenbeck, C. V. Hong, and J. I. Gordon, "Angiogenins: a new class of microbicidal proteins involved in innate immunity," *Nature Immunology*, vol. 4, no. 3, pp. 269–273, 2003.

[12] B. Rudolph, R. Podschun, H. Sahly, S. Schubert, J. M. Schröder, and J. Harder, "Identification of RNase 8 as a novel human antimicrobial protein," *Antimicrobial Agents and Chemotherapy*, vol. 50, no. 9, pp. 3194–3196, 2006.

[13] S. Lee-Huang, P. L. Huang, Y. Sun et al., "Lysozyme and RNases as anti-HIV components in β-core preparations of human chorionic gonadotropin," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 6, pp. 2678–2681, 1999.

[14] C. Sorkness, K. McGill, and W. W. Busse, "Evaluation of serum eosinophil cationic protein as a predictive marker for asthma exacerbation in patients with persistent disease," *Clinical and Experimental Allergy*, vol. 32, no. 9, pp. 1355–1359, 2002.

[15] M. Torrent, B. G. de la Torre, V. M. Nogués, D. Andreu, and E. Boix, "Bactericidal and membrane disruption activities of the eosinophil cationic protein are largely retained in an N-terminal fragment," *Biochemical Journal*, vol. 421, no. 3, pp. 425–434, 2009.

[16] M. Torrent, M. Badia, M. Moussaoui, D. Sanchez, M. V. Nogués, and E. Boix, "Comparison of human RNase 3 and RNase 7 bactericidal action at the gram-negative and gram-positive bacterial cell wall," *FEBS Journal*, vol. 277, no. 7, pp. 1713–1725, 2010.

[17] S. Navarro, J. Aleu, M. Jiménez, E. Boix, C. M. Cuchillo, and M. V. Nogués, "The cytotoxicity of eosinophil cationic protein/ribonuclease 3 on eukaryotic cell lines takes place through its aggregation on the cell membrane," *Cellular and Molecular Life Sciences*, vol. 65, no. 2, pp. 324–337, 2008.

[18] M. F. García-Mayoral, M. Moussaoui, B. G. de la Torre et al., "NMR structural determinants of eosinophil cationic protein binding to membrane and heparin mimetics," *Biophysical Journal*, vol. 98, no. 11, pp. 2702–2711, 2010.

[19] M. Torrent, S. Navarro, M. Moussaoui, M. V. Nogués, and E. Boix, "Eosinophil cationic protein high-affinity binding to bacteria-wall lipopolysaccharides and peptidoglycans," *Biochemistry*, vol. 47, no. 11, pp. 3544–3555, 2008.

[20] D. Pulido, M. Moussaoui, D. Andreu, M. V. Nogues, M. Torrent, and E. Boix, "Antimicrobial action and cell agglutination by the eosinophil cationic protein are modulated by the cell wall lipopolysaccharide structure," *Antimicrobial Agents and Chemotherapy*, vol. 56, no. 5, pp. 2378–2385, 2012.

[21] A. Trulson, J. Byström, A. Engström, R. Larsson, and P. Venge, "The functional heterogeneity of eosinophil cationic protein is determined by a gene polymorphism and post-translational modifications," *Clinical and Experimental Allergy*, vol. 37, no. 2, pp. 208–218, 2007.

[22] R. Sugihara, T. Kumamoto, T. Ito, H. Ueyama, I. Toyoshima, and T. Tsuda, "Human muscle protein degradation *in vitro* by eosinophil cationic protein (ECP)," *Muscle and Nerve*, vol. 24, no. 12, pp. 1627–1634, 2001.

[23] M. C. Pereira, D. T. Oliveira, and L. P. Kowalski, "The role of eosinophils and eosinophil cationic protein in oral cancer: a review," *Archives of Oral Biology*, vol. 56, no. 4, pp. 353–358, 2011.

[24] L. M. Zheutlin, S. J. Ackerman, G. J. Gleich, and L. L. Thomas, "Stimulation of basophil and rat mast cell histamine release

by eosinophil granule-derived cationic protein," *Journal of Immunology*, vol. 133, no. 4, pp. 2180–2185, 1984.

[25] J. B. Domachowske, K. D. Dyer, A. G. Adams, T. L. Leto, and H. F. Rosenberg, "Eosinophil cationic protein/RNase 3 is another RNase A-family ribonuclease with direct antiviral activity," *Nucleic Acids Research*, vol. 26, no. 14, pp. 3358–3363, 1998.

[26] K. C. Chang, C. W. Lo, T. C. Fan et al., "TNF-α mediates eosinophil cationic protein-induced apoptosis in BEAS-2B cells," *BMC Cell Biology*, vol. 11, article 6, 2010.

[27] T. C. Fan, S. L. Fang, C. S. Hwang et al., "Characterization of molecular interactions between eosinophil cationic protein and heparin," *The Journal of Biological Chemistry*, vol. 283, no. 37, pp. 25468–25474, 2008.

[28] T. C. Fan, H. T. Chang, I. W. Chen, H. Y. Wang, and M. D. T. Chang, "A heparan sulfate-facilitated and raft-dependent macropinocytosis of eosinophil cationic protein," *Traffic*, vol. 8, no. 12, pp. 1778–1795, 2007.

[29] R. J. Johnson, J. G. McCoy, C. A. Bingman, G. N. Phillips Jr., and R. T. Raines, "Inhibition of human pancreatic ribonuclease by the human ribonuclease inhibitor protein," *Journal of Molecular Biology*, vol. 368, no. 2, pp. 434–449, 2007.

[30] G. J. Gleich, D. A. Loegering, M. P. Bell, J. L. Checkel, S. J. Ackerman, and D. J. McKean, "Biochemical and functional similarities between human eosinophil-derived neurotoxin and eosinophil cationic protein: homology with ribonuclease," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 83, no. 10, pp. 3146–3150, 1986.

[31] J. E. Gabay, R. W. Scott, D. Campanelli et al., "Antibiotic proteins of human polymorphonuclear leukocytes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 86, no. 14, pp. 5610–5614, 1989.

[32] D. Yang, Q. Chen, H. F. Rosenberg et al., "Human ribonuclease A superfamily members, eosinophil-derived neurotoxin and pancreatic ribonuclease, induce dendritic cell maturation and activation," *Journal of Immunology*, vol. 173, no. 10, pp. 6134–6142, 2004.

[33] M. Seno, J. I. Futami, Y. Tsushima et al., "Molecular cloning and expression of human ribonuclease 4 cDNA," *Biochimica et Biophysica Acta*, vol. 1261, no. 3, pp. 424–426, 1995.

[34] R. D. I. Liddo, D. Dalzoppo, S. Baiguera et al., "*In vitro* biological activity of bovine milk ribonuclease-4," *Molecular Medicine Reports*, vol. 3, no. 1, pp. 127–132, 2010.

[35] S. K. Saxena, S. M. Rybak, R. T. Davey Jr., R. J. Youle, and E. J. Ackerman, "Angiogenin is a cytotoxic, tRNA-specific ribonuclease in the RNase A superfamily," *The Journal of Biological Chemistry*, vol. 267, no. 30, pp. 21982–21986, 1992.

[36] G. Tsirakis, C. A. Pappa, P. Kanellou et al., "Role of platelet-derived growth factor-AB in tumour growth and angiogenesis in relation with other angiogenic cytokines in multiple myeloma," *Hematological Oncology*, vol. 30, no. 3, pp. 131–136, 2012.

[37] H. F. Rosenberg and K. D. Dyer, "Molecular cloning and characterization of a novel human ribonuclease (RNase k6): increasing diversity in the enlarging ribonuclease gene family," *Nucleic Acids Research*, vol. 24, no. 18, pp. 3507–3513, 1996.

[38] M. Torrent, D. Sánchez, V. Buzón, M. V. Nogués, J. Cladera, and E. Boix, "Comparison of the membrane interaction mechanism of two antimicrobial RNases: RNase 3/ECP and RNase 7," *Biochimica et Biophysica Acta*, vol. 1788, no. 5, pp. 1116–1125, 2009.

[39] J. Harder and J. M. Schröder, "RNase 7, a novel innate immune defense antimicrobial protein of healthy human skin," *The Journal of Biological Chemistry*, vol. 277, no. 48, pp. 46779–46784, 2002.

[40] Y. C. Huang, Y. M. Lin, T. W. Chang et al., "The flexible and clustered lysine residues of human ribonuclease 7 are critical for membrane permeability and antimicrobial activity," *The Journal of Biological Chemistry*, vol. 282, no. 7, pp. 4626–4633, 2007.

[41] J. Zhang, K. D. Dyer, and H. F. Rosenberg, "RNase 8, a novel RNase A superfamily ribonuclease expressed uniquely in placenta," *Nucleic Acids Research*, vol. 30, no. 5, pp. 1169–1175, 2002.

[42] T. W. Pai, M. D. T. Chang, W. S. Tzou et al., "REMUS: a tool for identification of unique peptide segments as epitopes," *Nucleic Acids Research*, vol. 34, pp. W198–W201, 2006.

[43] E. Boix, Z. Nikolovski, G. P. Moiseyev, H. F. Rosenberg, C. M. Cuchillo, and M. V. Nogués, "Kinetic and product distribution analysis of human eosinophil cationic protein indicates a subsite arrangement that favors exonuclease-type activity," *The Journal of Biological Chemistry*, vol. 274, no. 22, pp. 15605–15614, 1999.

[44] A. Calabro, M. Benavides, M. Tammi, V. C. Hascall, and R. J. Midura, "Microanalysis of enzyme digests of hyaluronan and chondroitin/dermatan sulfate by fluorophore-assisted carbohydrate electrophoresis (FACE)," *Glycobiology*, vol. 10, no. 3, pp. 273–281, 2000.

[45] J. D. E. Young, C. G. B. Peterson, P. Venge, and Z. A. Cohn, "Mechanims of membrane damage mediated by human eosinophil cationic protein," *Nature*, vol. 321, no. 6070, pp. 613–616, 1986.

[46] I. D'Angelo, S. Welti, F. Bonneau, and K. Scheffzek, "A novel bipartite phospholipid-binding module in the neurofibromatosis type 1 protein," *EMBO Reports*, vol. 7, no. 2, pp. 174–179, 2006.

[47] I. Ishizuka, "Chemistry and functional distribution of sulfoglycolipids," *Progress in Lipid Research*, vol. 36, no. 4, pp. 245–319, 1997.

[48] S. C. Hung, X. A. Lu, J. C. Lee et al., "Synthesis of heparin oligosaccharides and their interaction with eosinophil-derived neurotoxin," *Organic and Biomolecular Chemistry*, vol. 10, no. 4, pp. 760–772, 2012.

[49] F. Soncin, D. J. Strydom, and R. Shapiro, "Interaction of heparin with human angiogenin," *The Journal of Biological Chemistry*, vol. 272, no. 15, pp. 9818–9824, 1997.

[50] E. Carreras, E. Boix, H. F. Rosenberg, C. M. Cuchillo, and M. V. Nogués, "Both aromatic and cationic residues contribute to the membrane-lytic and bactericidal activity of eosinophil cationic protein," *Biochemistry*, vol. 42, no. 22, pp. 6636–6644, 2003.

[51] M. M. Miyamoto, B. F. Koop, J. L. Slightom, M. Goodman, and M. R. Tennant, "Molecular systematics of higher primates: genealogical relations and classification," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 85, no. 20, pp. 7627–7631, 1988.

[52] H. F. Rosenberg, K. D. Dyer, H. L. Tiffany, and M. Gonzalez, "Rapid evolution of a unique family of primate ribonuclease genes," *Nature Genetics*, vol. 10, no. 2, pp. 219–223, 1995.

[53] G. Z. Cheng, J. Y. Li, F. Li, H. Y. Wang, and G. X. Shi, "Human ribonuclease 9, a member of ribonuclease A superfamily, specifically expressed in epididymis, is a novel sperm-binding protein," *Asian Journal of Andrology*, vol. 11, no. 2, pp. 240–251, 2009.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 410294, 9 pages
http://dx.doi.org/10.1155/2013/410294

Research Article

Using Nanoinformatics Methods for Automatically Identifying Relevant Nanotoxicology Entities from the Literature

Miguel García-Remesal,[1,2] Alejandro García-Ruiz,[2] David Pérez-Rey,[1,2] Diana de la Iglesia,[2] and Víctor Maojo[1,2]

[1] *Departamento de Inteligencia Artificial, Facultad de Informática, Universidad Politécnica de Madrid, Boadilla del Monte, 28660 Madrid, Spain*
[2] *Biomedical Informatics Group, Facultad de Informática, Universidad Politécnica de Madrid, Boadilla del Monte, 28660 Madrid, Spain*

Correspondence should be addressed to Miguel García-Remesal; mgarcia@infomed.dia.fi.upm.es

Received 8 May 2012; Revised 3 July 2012; Accepted 10 July 2012

Academic Editor: Raffaele Calogero

Nanoinformatics is an emerging research field that uses informatics techniques to collect, process, store, and retrieve data, information, and knowledge on nanoparticles, nanomaterials, and nanodevices and their potential applications in health care. In this paper, we have focused on the solutions that nanoinformatics can provide to facilitate nanotoxicology research. For this, we have taken a computational approach to automatically recognize and extract nanotoxicology-related entities from the scientific literature. The desired entities belong to four different categories: nanoparticles, routes of exposure, toxic effects, and targets. The entity recognizer was trained using a corpus that we specifically created for this purpose and was validated by two nanomedicine/nanotoxicology experts. We evaluated the performance of our entity recognizer using 10-fold cross-validation. The precisions range from 87.6% (targets) to 93.0% (routes of exposure), while recall values range from 82.6% (routes of exposure) to 87.4% (toxic effects). These results prove the feasibility of using computational approaches to reliably perform different named entity recognition (NER)-dependent tasks, such as for instance augmented reading or semantic searches. This research is a "proof of concept" that can be expanded to stimulate further developments that could assist researchers in managing data, information, and knowledge at the nanolevel, thus accelerating research in nanomedicine.

1. Introduction

Nanoinformatics is a nascent research field at the intersection of several disciplines, including informatics (information technologies and computer science), nanotechnology, medicine, biology, chemistry, and physics [1]. Nanoinformatics refers to theyu practical application of information technologies to gather, store, retrieve, and process information, data, and knowledge on the physicochemical characteristics of nanoparticles, nanomaterials, and nanodevices and their potential applications, especially in the biomedical field [1].

Applications of nanoinformatics include, for instance, nanoparticle characterization and design, modeling and simulation, data integration and exchange, linking nanoparticles information to clinical data, semantic annotation and retrieval, domain ontologies, terminologies and standards, and data and text mining for nanomedical research [2]. In this context, we can recall and emphasize the role that bioinformatics—a related informatics discipline—played in accelerating the Human Genome Project. One can conjecture that nanoinformatics might play the same role for nanotechnology and nanomedicine that bioinformatics and medical informatics have played in biology and medicine. We have

already begun to define the role that nanoinformatics could play for nanomedicine, as reported elsewhere [3, 4].

In our recent work in this field, we have focused on the challenges, opportunities, and solutions that nanoinformatics can provide to a critical subfield of nanomedicine: nanotoxicology. This discipline aims to determine whether and to what extent the unique properties of nanoparticles (that arise due to questions such as quantum size effects and/or their large surface-to-volume ratio) may present potential or real threats to humans, the environment, or to other species.

Publications have recently highlighted how nanoparticles enable a wide range of applications for clinical and therapeutic purposes. Bolhassani and colleagues [5] discuss the use of different types of nanoparticles, such as dendrimers, polymeric nanoparticles, metallic and magnetic nanoparticles, and quantum dots as effective vaccine adjuvants for infectious diseases and cancer therapy. Kosuge and colleagues [6] report the use of FeCo/graphitic-carbon nanocrystals (FeCo/GCNs) to enhance cellular fluorescence and magnetic resonance imaging of vascular inflammation due to their accumulation in vascular macrophages in vivo. Similarly, Thakor and colleagues [7] describe the use of polyethylene glycosylated Raman-active gold nanoparticles (PEG-R-AuNPs) in different clinical trials targeting dysplastic bowel lesions during colonoscopy. More extensive reviews can be found in [8, 9].

Despite these advances, the use of nanoparticles may involve serious risks for both patients and environment due to potential secondary toxic effects, also reported in the literature [10–14]. Therefore, it is essential for clinicians and researchers using nanoparticles for therapeutic purposes to be able to access relevant nanotoxicology information in an integrated and intuitive manner. Similarly, regulatory and environmental researchers need data and information integration in performing risk assessments or environmental forecasts as the result of manufacturing, use, degradation, disposal, and recycling of these materials. Taking advantage of nanoinformatics methods—most specifically text mining and natural language processing techniques applied to toxicological issues—should contribute to automatically identifying, organizing and making available specific nanotoxicity information reported in the literature to researchers and physicians.

Based on related research, we have carried out in the Biomedical Informatics field (BMI) [15–18], we present in this paper a nanoinformatics approach based on named entity recognition (NER) techniques for automatically extracting nanotoxicology-related entities from the literature. This, to our knowledge, is the first reported effort to automatically identify and extract relevant entities from scientific papers relevant to nanotoxicology. The extracted entities include, for instance, names of nanoparticles, nanomaterials, and nanodevices, types of toxicity/damage—for example, cell death or lung inflammation—and potential routes of exposure to toxic agents—for instance, inhalation or dermal contact. Once this information is retrieved and gathered, it can be used for a wide variety of applications.

This paper is organized as follows. In the background Section 2, we provide a survey of existing NER-focused methods and tools, most of them developed in the context of bioinformatics and medical informatics research. In the methods Section 3, we describe the building of the nanotoxicology training corpus, the training and construction of the automated entity recognizer, and the design of the evaluation experiment. Next, we present and discuss the results of the evaluation. Finally, we present the conclusions.

2. Background

Over the past few years, named Entity recognition (NER) methods and techniques have been widely used in medical informatics and bioinformatics research to automatically identify and extract different types of named entities (NEs) such as gene and/or protein names [19–23], medications and dosages [24], primary diseases and comorbidities [24], or raw sequences of nucleic acids and proteins [16, 20, 25, 26].

According to Park and Kim [27], there are four main approaches to performing NER from textual sources: (1) dictionary-based approaches, (2) rule-based approaches, (3) machine-learning approaches and (4) hybrid approaches. Dictionary-based approaches, try to identify entity names belonging to domain-specific controlled vocabularies, taxonomies and/or ontologies directly from the literature. There are different techniques for matching entities mentioned in the text to dictionary entries. These include, simple pattern-matching [28–30] or statistical techniques [31] to compare sequences of tokens from the text to dictionary entries, advanced symbolic natural language processing and computational linguistic techniques such as those used in the National Library of Medicine's MetaMap program [32, 33], and innovative hybrid approaches such as the one described in [34]. This encodes both biomedical texts and dictionary entries into sequences of nucleotide symbols—i.e., A, C, G, and T. Once the dictionary entries and the textual documents have been converted into sequences, the authors use BLAST [35]—the most ubiquitous tool for DNA and protein sequence matching—to automatically identify the entity instances in the text. Although dictionary-based approaches are relatively simple to design and implement if the appropriate dictionary is available, they have several limitations. These include false positive and false negative recognition issues arising from ambiguous names and from synonym and spelling variants, respectively.

Rule-based approaches address some of the limitations of dictionary-based approaches by dealing with morphological variants not covered by the latter approaches [27]. Rule-based methods resort to handcrafted patterns and rules to deal with the different types of morphological variants. Some examples of rule-based approaches include [36–42]. The main disadvantage of rule-based approaches is the difficulty to adapt or reuse them for different domains.

In contrast to rule-based methods—that use handcrafted rules and patterns—machine learning approaches are aimed at "learning" predictive models that can be used to automatically detect the occurrence of NEs in the text. Examples of machine learning methods and techniques used for

NER include conditional random fields [21, 43–46], hidden markov models [47–49], support vector machines [45], and context-aware rule-based classifiers [33, 50]. To automatically generate the desired predictive models, nearly all machine learning-based approaches require a set of documents to train the model. This training set is a body of text documents (often just single passages) that has been manually analyzed and annotated by domain experts to identify different entities occurring in them. Examples of widely used corpora in the biological domain include GENIA [51]—an annotated body of literature related to the MeSH terms "human", "blood cells", and "transcription factors"—the BioCreAtIvE body for Task 1A [52]—text passages annotated with names of genes and related entities—, or Linnaeus [53]—aimed at recognizing and identifying species names in the biomedical literature. Similar corpora—although considerably smaller in size—have been developed for the medical domain. These include, for instance, the corpora used in the I2B2 medication extraction challenge [24, 54] and the I2B2 Obesity NLP Challenge [55], or a recently developed corpus aimed at the automated discovery of anaphoric relations in clinical narrative [56].

Finally, hybrid approaches combine two or more of the previously described techniques to achieve better performance, since each of the described approaches have its own advantages and disadvantages. Examples of hybrid systems approach include [45], which combines two machine learning algorithms (conditional random fields and support vector machines) with several rule-based engines, the approaches described in [16, 33], that rely on rule-based systems and lookup lists, or the hybrid method reported in [20] that describes a system combining a preprocessing dictionary and a rule-based filter with several independently trained support vector machines.

After reviewing the results of recent NER-related challenges [24, 55], we decided to adopt a machine learning approach based on conditional random fields (CRFs) to build our nanotoxicology-related named entity recognizer. We made this decision since CRF-based biomedical NER systems are fast, effective, accurate, and perform relatively well even if trained with small training sets [21, 24, 55]. The latter issue is critical for the purpose reported in this paper, since to our knowledge there are no any available corpora for the nanotoxicological domain. Therefore, we had to build our own nanotoxicology corpus from scratch, which is a difficult and time-consuming task.

In the next section, we describe (1) the methods we used to build the corpus for training and evaluating the recognizer, (2) the CRF training process, and (3) the metrics we used to evaluate the performance of the nanotoxicology-related named entity recognizer.

3. Methods

The proposed NER system is designed to recognize instances of entities belonging to four different categories: NANO, EXPO, TOXIC, and TARGET. Entities belonging to the NANO category represent nanoparticles, nanodevices, or nanomaterials, such as for instance, *"polyamidoamine dendrimers"* or *"buckminsterfullerene"*. Similarly, EXPO-labeled instances describe different routes of human, animal, or environmental exposure to nanoparticles, such as *"inhalation"*, *"dermal contact"* or *"pulverization"*. On the other hand, TOXIC-labeled terms represent toxicological hazards of nanoparticles such as *"detachment"* or *"death"*, while TARGET-labeled terms refer to the actual targets of the hazards such as *"cell"* or *"kidney"*.

The purpose of this study was to review published dose-response data on acute lung inflammation in rats after instillation of titanium dioxide particles or six types of carbon nanoparticles.

The purpose of this study was to review published dose-response data on acute <TARGET> lung </TARGET> <TOXIC> inflammation </TOXIC> in <TARGET> rats </TARGET> after <EXPO> instillation </EXPO> of <NANO> titanium dioxide particles </NANO> or six types of <NANO> carbon nanoparticles </NANO>

Figure 1: Sample annotated sentence belonging to the current "gold standard", containing 6 different mentions of entities belonging to different categories.

We trained a CRF model using an annotated corpus containing 300 sentences selected from the available literature. Further details on the creation of the annotated corpus, the training of the CRF model and the evaluation protocol follow.

3.1. Building the Annotated Corpus. To build the corpus, we submitted the query "nanoparticles/toxicity(MeSH major topic)" to PubMed, obtaining 654 results at the time of writing. We manually analyzed the resulting set of abstracts to choose 300 sentences containing relevant entities. Members of our research group manually annotated the selected sentences. Both the selection of the 300 sentences and the annotation process were validated by two experts in nanomedicine and nanotoxicology.

The outcome of the labeling process was an annotated set of 300 sentences. Figure 1 shows a sample annotated sentence containing instances for all the target categories. As depicted in the figure, each entity is enclosed between an opening and ending tag that denotes the category to which it belongs. For this sample sentence, we have two different instances belonging to the NANO category: *"titanium dioxide particles"* and *"carbon nanoparticles"*, one to EXPO: *"instillation"*, two

Table 1: Number of entities and tokens manually identified by the annotators in the 300 selected phrases and annotated as belonging to one of the target categories.

	Nano	Expo	Toxic	Target	Total
Entities	426	144	485	385	1440
Tokens	717	186	637	705	2245

to TARGET: *"lung"* and *"rats"*, and one to TOXIC: *"inflammation"*.

Table 1 summarizes the number of entities belonging to each category that were identified in the 300 selected phrases and labeled as such by the annotators. As entities may be composed of 2 or more words (tokens), such as for instance *"titanium dioxide particles"*, the table also reports the total number of tokens belonging to each category. Thus, the mention *"titanium dioxide particles"*, belonging to the NANO category at entity-level would be counted as 3 different mentions of the NANO category at the token-level. We made this distinction to evaluate the performance of the system both at entity-level (exact matching) and at token level (partial or inexact matching). Further detail is given in the section *Evaluation Metrics*.

3.2. Training the CRF Model. We trained a CRF model on the 300 annotated sentences to automatically identify instances of entities belonging to the four target categories. To train the CRF, we used the Java Application Programming Interface (API) provided by ABNER [21]. The latter is an open-source named entity recognizer designed to identify protein names and gene products. The model was trained using the default set of features provided by ABNER that includes orthographic, morphological, and contextual features. The latter are mostly based on regular expressions and n-gram features. We also performed minor modifications on the default tokenizer supplied with ABNER to properly identify chemical formulas.

3.3. Evaluation Metrics. We assessed the performance of the CRF-based NER system by calculating the precision, recall, and *F*-measure values for each type of entity—that is, NANO, EXPO, TOXIC, and TARGET. These metrics were computed both at entity and token levels [54]. Entity-level metrics measure the ability of the system to successfully recognize the full text of multiword entities labeled as such in the gold standard—i.e., the training set of manual annotations in the corpus. Conversely, token-level metrics are targeted at evaluating the performance of the system when labeling individual words. For instance, let us suppose that the annotation provided by our system for the sentence *"In this study, metallic nickel nanoparticles caused higher…"* is *"In this study, metallic <NANO>nickel nanoparticles</NANO> caused higher…"*, and that the provided annotation for this sample sentence in the gold standard is *"In this study, <NANO>metallic nickel nanoparticles</NANO> caused higher…"*. Therefore, for this example, the system would fail to provide a correct entity-level annotation for the NANO-labeled entity "*metallic nickel nanoparticles*", since the system only achieved a partial match. However, this annotation would lead to an increase in recall for the NANO category at the token level, since the system successfully recognized 2 tokens (out of 3) in the phrase "*metallic nickel nanoparticles*" as belonging to the NANO category. We used formulas (1) to compute entity-level and token-level precision, recall, and *F*-measure:

$$\begin{aligned}
\text{Entity-level Precision (EP)} &= \frac{\#\text{correctly returned entities by system}}{\#\text{entities returned by system}},\\
\text{Entity-level Recall (ER)} &= \frac{\#\text{correctly returned entities by system}}{\#\text{entities in gold standard}},\\
\text{Entity-level } F\text{-measure (EF)} &= \frac{2\cdot \text{EP}\cdot \text{ER}}{\text{EP}+\text{ER}},\\
\text{Token-level Precision (TP)} &= \frac{\#\text{correctly returned tokens from each entity in system output}}{\#\text{ tokens in system output}},\\
\text{Token-level Recall (TR)} &= \frac{\#\text{correctly returnedtokens from each entity in system output}}{\#\text{ tokens in gold standard}},\\
\text{Token-level } F\text{-measure (EF)} &= \frac{2\cdot \text{TP}\cdot \text{TR}}{\text{TP}+\text{TR}}.
\end{aligned} \tag{1}$$

Although the size of the set of annotated sentences—in terms of number of sentences, entities, and tokens—is reasonable and could be divided into a training and test set to evaluate the system's performance, we instead chose to use 10-fold cross-validation to avoid overfitting. In the next section, we report the results of the evaluation activity.

4. Results and Discussion

Table 2 summarizes the results of the evaluation of the CRF-based entity identifier against the manually annotated gold standard using 10-fold cross-validation. The table shows the precision, recall, and *F*-measure for each target category

Table 2: Summary of results of the evaluation of the CRF-based recognizer using 10-fold cross-validation.

	Entity-level			Token-level		
	Precision (EP)	Recall (ER)	*F*-measure (FR)	Precision (TP)	Recall (TR)	*F*-measure (TF)
Nano	0.892	0.873	0.883	0.945	0.943	0.944
Expo	0.930	0.826	0.875	0.981	0.855	0.914
Toxic	0.926	0.874	0.899	0.967	0.909	0.937
Target	0.876	0.860	0.868	0.906	0.916	0.911

(NANO, EXPO, TOXIC, and TARGET) both at entity and token level.

As shown in Table 2, our CRF-based entity recognizer yields entity-level precision values that range from 87.60% (TARGET) to 93.00% (EXPO). Similarly, entity-level recall values range from 82.60% (EXPO) to 87.40% (TOXIC). Performance of the recognizer at the token level, include precision values ranging from 90.06% (TARGET) to 98.10% (EXPO), while recall values range from 85.50% (EXPO) to 94.30% (NANO). These results show that the CRF-based approach performs particularly well at recognizing nanotoxicology entities—with entity-level *F*-measure values close to 90% for all categories—even when trained with such a reduced set of sentences. Moreover, the CRF-based approach seems to perform better at recognizing nanotoxicology entities than at identifying entities belonging to the biomedical domain, as reported for protein and gene names (precision = 65.90%, recall = 74.50%) [21], or medication information (precision = 90.37%, recall = 66.12%) [44]. The targeted entities are, of course, quite different, so direct comparisons should be treated with caution.

To ensure a fair evaluation, we compared the adopted CRF-based approach to a hybrid approach used as baseline. This hybrid method combines a dictionary-based approach with a term selection scheme based on TF/IDF (term frequency/inverse document frequency) weights [57]. The latter are widely accepted statistics that measure the importance of a term in the context of a textual collection or corpus. To evaluate the hybrid method used as baseline, we proceeded as follows. First, we built a dictionary containing all terms occurring in the corpus we created, composed of 300 sentences. This dictionary contained all tokens—excluding stop words—and n-grams of sizes ranging between 2 and 6 occurring in the corpus. N-grams are groups of tokens that appear consecutively in the text. For instance, for the sentence *"Gold nanoparticles have the potential to …"* we would have the following n-grams of size 2: *"Gold nanoparticles"*, *"nanoparticles have"*, *"have the"*, *"the potential"*, *"potential to"*. Examples of n-grams of size 3 include *"Gold nanoparticles have"* or *"nanoparticles have the"*. We chose using n-grams in addition to single-word tokens since many concepts belonging to different ontologies are multiword concepts. Next, for each term T in the vocabulary, we calculated its TF/IDF score for the document containing the maximum number of occurrences of the term T. After that, all terms in the vocabulary were sorted in descending order of the TF/IDF score. We discarded all terms having a TF/IDF score smaller than *0.1*. Finally, we compared the remaining terms in the vocabulary to terms belonging to two different ontologies: the Foundational Model of Anatomy [58]—to detect anatomical locations that might be potential targets of nanoparticles—and the Nanoparticle Ontology [59]—to identify names of nanoparticles. If a term from the vocabulary matched a term from any of the ontologies, then it was marked as belonging to the NANO category—if the matched term belonged to the Nanoparticle Ontology—or to the TARGET category—if the matched term belonged to the Foundational Model of Anatomy. Note that, we did not focus on identifying toxic effects of nanoparticles and modes of exposition since there are no currently available ontologies or controlled vocabularies addressing such topics, and thus it is not possible using a vocabulary-based approach. Table 3 shows the results of the evaluation for the method used as baseline.

Table 3: Summary of results of the evaluation of the hybrid approach used as baseline.

	Entity-level		
	Precision	Recall	*F*-measure
Nano	1.00	0.33	0.496
Target	0.75	0.48	0.585

As shown in Table 3, the baseline approach yields precisions of 100% and 75% for the NANO and TARGET categories respectively. These figures are reasonable, since most terms matching concepts belonging to the Nanoparticle Ontology refer to names of nanoparticles with high probability. This is not the case, however, for terms matching concepts from the Foundational Model of Anatomy, since anatomical locations may be mentioned together with nanoparticle names and there not might exist any toxicity relationships between them. Regarding the recall values yielded by the baseline method, it must be noted that these values are much smaller than those yielded by the CRF-based approach. These values are also reasonable, since the Nanoparticle Ontology was initially designed to provide a conceptualization of the domain of cancer nanotechnology research, while the documents in the corpus are targeted at different diseases. Similarly, the Foundational Model of Anatomy alone is not suitable for detecting potential targets of nanoparticles, since in addition to anatomical locations, potential targets of nanoparticles may also include animals and the environment.

These results suggest that the CRF-based approach is suitable for performing NER-dependent tasks, especially when other approaches such as the vocabulary-based one

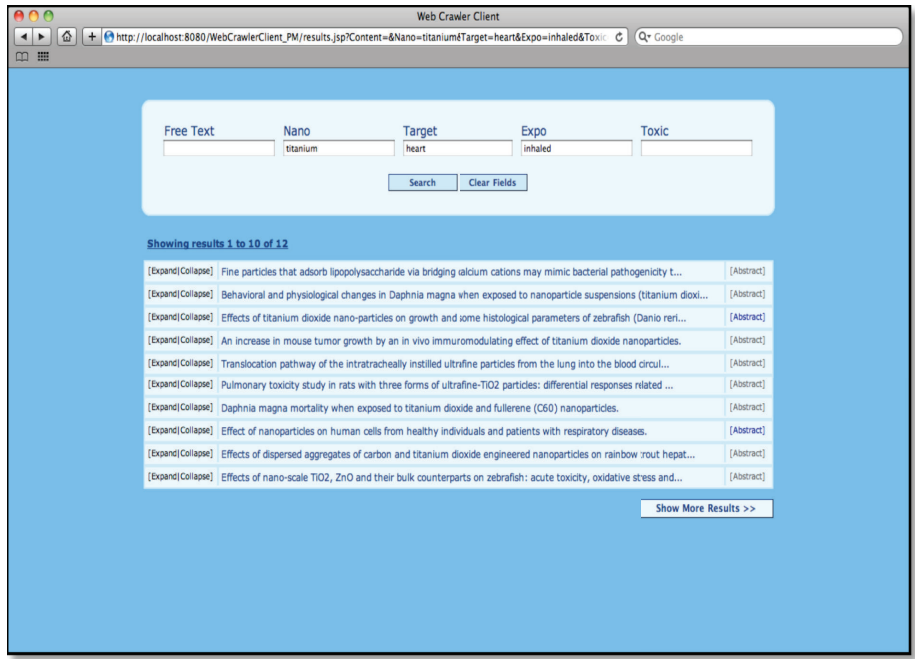


Figure 2: Screenshot of the prototype of the "nanotoxicity searcher".

cannot be performed due to the lack of a well-established controlled vocabularies or ontologies.

Examples of NER-dependant tasks that can be carried out using our nanotoxicology recognizer include, for instance, finding relationships between the detected entities, or indexing scientific papers with the different entities appearing in them. In fact, the latter task is a significant research topic in biomedical informatics research, since many different systems for automatically indexing and searching the biomedical literature have been developed over the last few years. Examples include Pharmspresso [60], an information retrieval and extraction system for pharmacogenomic-related literature that follows a dictionary-based approach to identify instances of genes, drugs, polymorphisms and diseases, or PubDNA Finder [17], an online repository that we developed to link PubMed Central manuscripts to the sequences of nucleic acids appearing in them, following a hybrid approach that combines a rule-based system and lookup lists. We have already begun working in this direction with the development of a prototype of the "nanotoxicity searcher". The latter is an intelligent search engine that provides users with a web interface to search for PubMed-indexed papers that were automatically annotated with specific mentions of relevant nanotoxicology entities using the methods described in this paper. Figure 2 shows a screenshot of the current prototype of the "toxicity searcher". We believe that our search engine can be a valuable tool for nanomedical researchers to easily discover toxic and secondary effects of nanoparticles reported in the literature.

To our knowledge, the results we report are the first application of text mining methods to extract nanotoxicology information from the literature—in fact, the first text mining application in the whole field of nanomedicine. Considerable interaction between nanoinformatics professionals should enable building extended corpora in this and other fields, where challenges and competitive testing can be carried out to evaluate these methods from text mining, information retrieval, and how they perform with different information types. Similar competitions have been earlier carried out in BMI, with significant results and success [52, 54, 55]. In this way, we can consider our research as a first "proof of concept", which needs to be followed up soon by efforts by others, and may provide opportunities in an entirely new area of research for nanoinformaticians.

Extending the research presented in this work to include more general entities—that is, nanomedicine and nanotechnology-related entities—can open new and significant challenges for nanoinformaticians, given the novelty of this topic and approach. These potential challen-ges include, for instance: (a) populating electronic health records and/or clinical trials with nanolevel information extracted from the literature, (b) automatically annotating and indexing nanomedical documents mentioning concepts and entities belonging to well-known ontologies and controlled vocabularies, (c) aligning and bringing together existing biomedical and nanomedicine/nanotechnology ontologies—such as for example the Nanoparticle Ontology [59], the Foundational Model of Anatomy [58], or the Gene Ontology [61]—or (d) automatically creating inventories of nanoparticles containing details about their characterization and design, potential uses, and applications—for example tissue regeneration, drug delivery, medical imaging, identification of cancerous cells, for example—toxicity, links to relevant literature, links to modeling, and simulation tools, and so forth.

This research is an example of the potential challenges and synergies that lie ahead for future interactions between experts in nanotechnology, nanomedicine, and

nanoinformatics. Such interactions may lead to a broad range of medical applications involving different nanomedical challenges. In this regard, the authors are currently working together on the development of new methods and tools for addressing these issues.

5. Conclusion

In this paper, we have presented a nanoinformatics approach based on NER techniques for automatically identifying relevant nanotoxicology entities in scientific articles. The results of the evaluation suggest that the entity recognizer, we have developed could be used by other nanoinformaticians to reliably perform different NER-dependant tasks. These include extracting nanotoxicity information from textual sources to populate structured databases, or to automatically index and search nanotoxicology articles. In addition, this work can be extended to recognizing more general nanomedicine and nanotechnology entities, thus providing new research opportunities for nanoinformaticians. This is, to our knowledge, the first report that explores the use of text mining techniques in the area of nanotechnology. Further research in this emerging nanoinformatics field may lead to the development of novel methods and tools that could assist researchers in managing data, information, and knowledge at the nanolevel, thus accelerating research in nanoscience.

Conflict of Interests

The authors declare that they have no conflict of interests.

Acknowledgments

This research has been partially funded by the European Commission (the ACTION-Grid Support Action, FP7-224176), the Spanish Ministry of Economy and Competitiveness (FIS/AES PS09/00069, RETICS COMBIOMED RD07/0067/0006, Ibero-NBIC CYTED 209RT0366), the "Consejo Social of the Universidad Politécnica de Madrid", and the "Comunidad de Madrid". The authors would also like to thank Professor Casimir A. Kulikowski, Dr. Martin Fritts, and Dr. Raul E. Cachau for their useful suggestions and comments.

References

[1] V. Maojo, F. Martin-Sanchez, C. Kulikowski, A. Rodriguez-Paton, and M. Fritts, "Nanoinformatics and DNA-based computing: catalyzing nanomedicine," *Pediatric Research*, vol. 67, no. 5, pp. 481–489, 2010.

[2] V. Maojo, M. García-Remesal, D. de la Iglesia, and J. Crespo, "Nanoinformatics: developing advanced informatics applications for Nanomedicine," in *Intracellular Drug Delivery: Fundamentals and Applications*, A. Prokop, Ed., vol. 5, pp. 847–860, 2011.

[3] ACTION-Grid consortium, The ACTION-Grid White Paper on Nanoinformatics, http://www.action-grid.eu/index.php?url=whitepaper, 2010.

[4] D. de la Iglesia, V. Maojo, S. Chiesa et al., "International efforts in nanoinformatics research applied to nanomedicine.," *Methods of Information in Medicine*, vol. 50, no. 1, pp. 84–95, 2011.

[5] A. Bolhassani, S. Safaiyan, and S. Rafati, "Improvement of different vaccine delivery systems for cancer therapy," *Molecular Cancer*, vol. 10, p. 3, 2011.

[6] H. Kosuge, S. P. Sherlock, T. Kitagawa et al., "FeCo/graphite nanocrystals for multi-modality imaging of experimental vascular inflammation," *PLoS ONE*, vol. 6, no. 1, Article ID e14523, 2011.

[7] A. S. Thakor, R. Paulmurugan, P. Kempen et al., "Oxidative stress mediates the effects of raman-active gold nanoparticles in human cells," *Small*, vol. 7, no. 1, pp. 126–136, 2011.

[8] R. A. Freitas, *Nanomedicine, Volume I: Basic Capabilities*, Landes Bioscience, Georgetown, Tex, USA, 1999.

[9] R. A. Freitas, *Nanomedicine, Volume IIA: Biocompatibility*, Landes Bioscience, Georgetown, Tex, USA, 2003.

[10] M. Li, L. Zhu, and D. Lin, "Toxicity of ZnO nanoparticles to Escherichia coli: mechanism and the influence of medium components," *Environmental Science and Technology*, vol. 45, no. 5, pp. 1977–1983, 2011.

[11] R. Hu, L. Zheng, T. Zhang et al., "Mechanism of inflammatory responses in brain and impairment of spatial memory of mice caused by titanium dioxide nanoparticles," *Journal of Hazardous Materials*, 2011.

[12] J. Chen, X. Dong, Y. Xin, and M. Zhao, "Effects of titanium dioxide nano-particles on growth and some histological parameters of zebrafish (Danio rerio) after a long-term exposure," *Aquatic Toxicology*, vol. 101, no. 3-4, pp. 493–499, 2011.

[13] A. Marushima, K. Suzuki, Y. Nagasaki et al., "Newly synthesized radical-containing nanoparticles enhance neuroprotection after cerebral ischemia-reperfusion injury," *Neurosurgery*, vol. 68, no. 5, pp. 1418–1425, 2011.

[14] A. Sharma, A. Tandon, J. C. K. Tovey et al., "Polyethylenimine-conjugated gold nanoparticles: gene transfer potential and low toxicity in the cornea," *Nanomedicine*, vol. 7, no. 4, pp. 505–513, 2011.

[15] M. García-Remesal, V. Maojo, J. Crespo, and H. Billhardt, "Logical schema acquisition from text-based sources for structured and non-structured biomedical sources integration," *AMIA Annual Symposium Proceedings*, pp. 259–263, 2007.

[16] M. García-Remesal, A. Cuevas, V. López-Alonso et al., "A method for automatically extracting infectious disease-related primers and probes from the literature," *BMC Bioinformatics*, vol. 11, p. 410, 2010.

[17] M. García-Remesal, A. Cuevas, D. Pérez-Rey et al., "PubDNA Finder: a web database linking full-text articles to sequences of nucleic acids," *Bioinformatics*, vol. 26, no. 21, Article ID btq520, pp. 2801–2802, 2010.

[18] M. García-Remesal, V. Maojo, H. Billhardt, and J. Crespo, "Integration of relational and textual biomedical sources: a pilot experiment using a semi-automated method for logical schema acquisition," *Methods of Information in Medicine*, vol. 49, no. 4, pp. 337–348, 2010.

[19] J. T. Chang, H. Schütze, and R. B. Altman, "GAPSCORE: finding gene and protein names one word at a time," *Bioinformatics*, vol. 20, no. 2, pp. 216–225, 2004.

[20] S. Mika and B. Rost, "NLProt: extracting protein names and sequences from papers," *Nucleic Acids Research*, vol. 32, supplement 2, pp. W634–W637, 2004.

[21] B. Settles, "ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text," *Bioinformatics*, vol. 21, no. 14, pp. 3191–3192, 2005.

[22] L. Tanabe, N. Xie, L. H. Thom, W. Matten, and W. J. Wilbur, "GENETAG: a tagged corpus for gene/protein named entity recognition," *BMC Bioinformatics*, vol. 6, supplement 1, p. S3, 2005.

[23] M. Torii, Z. Hu, C. H. Wu, and H. Liu, "BioTagger-GM: a gene/protein name recognition system," *Journal of the American Medical Informatics Association*, vol. 16, no. 2, pp. 247–255, 2009.

[24] Ö. Uzuner, I. Solti, and E. Cadag, "Extracting medication information from clinical text," *Journal of the American Medical Informatics Association*, vol. 17, no. 5, pp. 514–518, 2010.

[25] J. D. Wren, W. H. Hildebrand, S. Chandrasekaran, and U. Melcher, "Markov model recognition and classification of DNA/protein sequences within large text databases," *Bioinformatics*, vol. 21, no. 21, pp. 4046–4053, 2005.

[26] S. Aerts, M. Haeussler, S. van Vooren et al., "Text-mining assisted regulatory annotation," *Genome Biology*, vol. 9, no. 2, p. R31, 2008.

[27] J. C. Park and J. Kim, "Named entity recognition," in *Text Mining for Biology and Biomedicine*, S. Ananiadou and J. McNaught, Eds., pp. 121–142, Artech House, Norwood, Mass, USA, 2006.

[28] T. K. Jenssen, A. Lægreid, J. Komorowski, and E. Hovig, "A literature network of human genes for high-throughput analysis of gene expression," *Nature Genetics*, vol. 28, no. 1, pp. 21–28, 2001.

[29] T. Ono, H. Hishigaki, A. Tanigami, and T. Takagi, "Automated extraction of information on protein-protein interactions from the biological literature," *Bioinformatics*, vol. 17, no. 2, pp. 155–161, 2001.

[30] T. Hamon and N. Grabar, "Linguistic approach for identification of medication names and related information in clinical narratives," *Journal of the American Medical Informatics Association*, vol. 17, no. 5, pp. 549–554, 2010.

[31] R. Farkas, G. Szarvas, I. Hegedus et al., "Semi-automated Construction of Decision Rules to Predict Morbidities from Clinical Texts," *Journal of the American Medical Informatics Association*, vol. 16, no. 4, pp. 601–605, 2009.

[32] A. R. Aronson, "Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program," *Proceedings of the Annual Symposium Proceedings (AMIA '01)*, pp. 17–21, 2001.

[33] J. G. Mork, O. Bodenreider, D. Demner-Fushman et al., "Extracting Rx information from clinical narrative," *Journal of the American Medical Informatics Association*, vol. 17, no. 5, pp. 536–539, 2010.

[34] M. Krauthammer, A. Rzhetsky, P. Morozov, and C. Friedman, "Using BLAST for identifying gene and protein names in journal articles," *Gene*, vol. 259, no. 1-2, pp. 245–252, 2000.

[35] S. F. Altschul, T. L. Madden, A. A. Schäffer et al., "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, no. 17, pp. 3389–3402, 1997.

[36] K. Fukuda, A. Tamura, T. Tsunoda, and T. Takagi, "Toward information extraction: identifying protein names from biological papers.," *Pacific Symposium on Biocomputing*, pp. 707–718, 1998.

[37] D. Proux, F. Rechenmann, L. Julliard, V. V. Pillet, and B. Jacq, "Detecting gene symbols and names in biological texts: a first step toward pertinent information extraction," in *Proceedings of the Workshop on Genome Informatics*, vol. 9, pp. 72–80, 1998.

[38] R. Gaizauskas, G. Demetriou, and K. Humphreys, "Term recognition and classification in biological science journal articles," in *Proceedings of the Workshop on Computational Terminology for Medical and Biological Applications*, pp. 37–44, 2000.

[39] L. C. Childs, R. Enelow, L. Simonsen, N. H. Heintzelman, K. M. Kowalski, and R. J. Taylor, "Description of a Rule-based System for the i2b2 Challenge in Natural Language Processing for Clinical Data," *Journal of the American Medical Informatics Association*, vol. 16, no. 4, pp. 571–575, 2009.

[40] N. K. Mishra, D. M. Cummo, J. J. Arnzen, and J. Bonander, "A Rule-based Approach for Identifying Obesity and Its Comorbidities in Medical Discharge Summaries," *Journal of the American Medical Informatics Association*, vol. 16, no. 4, pp. 576–579, 2009.

[41] I. Spasić, F. Sarafraz, J. A. Keane, and G. Nenadić, "Medication information extraction with linguistic pattern matching and semantic rules," *Journal of the American Medical Informatics Association*, vol. 17, no. 5, pp. 532–535, 2010.

[42] H. Yang, "Automatic extraction of medication information from medical discharge summaries," *Journal of the American Medical Informatics Association*, vol. 17, no. 5, pp. 545–548, 2010.

[43] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: probabilistic models for segmenting and labeling sequence data," in *Proceedings of the 18th International Conference on Machine Learning*, pp. 282–289, Morgan Kaufmann, Williamstown, Mass, USA, 2001.

[44] Z. Li, F. Liu, L. Antieau, Y. Cao, and H. Yu, "Lancet: a high precision medication event extraction system for clinical text," *Journal of the American Medical Informatics Association*, vol. 17, no. 5, pp. 563–567, 2010.

[45] J. Patrick and M. Li, "High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge," *Journal of the American Medical Informatics Association*, vol. 17, no. 5, pp. 524–527, 2010.

[46] D. Tikk and I. Solt, "Improving textual medication extraction using combined conditional random fields and rule-based systems," *Journal of the American Medical Informatics Association*, vol. 17, no. 5, pp. 540–544, 2010.

[47] N. Collier, C. Nobata, and J. Tsujii, "Extracting the names of genes and gene products with a hidden Markov model," in *Proceedings of the 18th Conference on Computational Linguistics*, vol. 1, pp. 201–207, 2000.

[48] G. Zhou and J. Su, "Named entity recognition using an HMM-based chunk tagger," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pp. 473–480, 2002.

[49] A. A. Morgan, L. Hirschman, M. Colosimo, A. S. Yeh, and J. B. Colombe, "Gene name identification and normalization using a model organism database," *Journal of Biomedical Informatics*, vol. 37, no. 6, pp. 396–410, 2004.

[50] I. Solt, D. Tikk, V. Gál, and Z. T. Kardkovács, "Semantic classification of diseases in discharge summaries using a context-aware rule-based classifier," *Journal of the American Medical Informatics Association*, vol. 16, no. 4, pp. 580–584, 2009.

[51] T. Ohta, Y. Tateisi, J. D. Kim, and J. Tsujii, "The GENIA Corpus: an annotated corpus in molecular biology domain," in *Proceedings of the 2nd International Conference on Human Language Technology Research*, pp. 82–86, 2002.

[52] A. Yeh, A. Morgan, M. Colosimo, and L. Hirschman, "BioCreAtIvE task 1A: gene mention finding evaluation," *BMC Bioinformatics*, vol. 6, supplement 1, p. S2, 2005.

[53] M. Gerner, G. Nenadic, and C. M. Bergman, "LINNAEUS: a species name identification system for biomedical literature," *BMC Bioinformatics*, vol. 11, p. 85, 2010.

[54] Ö. Uzuner, I. Solti, F. Xia, and E. Cadag, "Community annotation experiment for ground truth generation for the i2b2 medication challenge," *Journal of the American Medical Informatics Association*, vol. 17, no. 5, pp. 519–523, 2010.

[55] Ö. Uzuner, "Recognizing Obesity and Comorbidities in Sparse Data," *Journal of the American Medical Informatics Association*, vol. 16, no. 4, pp. 561–570, 2009.

[56] G. K. Savova, W. W. Chapman, J. Zheng, and R. S. Crowley, "Anaphoric relations in the clinical narrative: corpus creation," *Journal of the American Medical Informatics Association*, vol. 18, no. 4, pp. 459–465, 2011.

[57] K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of Documentation*, vol. 60, no. 5, pp. 493–502, 2004.

[58] C. Rosse and J. L. V. Mejino Jr, "A reference ontology for biomedical informatics: the foundational model of anatomy," *Journal of Biomedical Informatics*, vol. 36, no. 6, pp. 478–500, 2003.

[59] D. G. Thomas, R. V. Pappu, and N. A. Baker, "NanoParticle ontology for cancer nanotechnology research," *Journal of Biomedical Informatics*, vol. 44, no. 1, pp. 59–74, 2011.

[60] Y. Garten and R. B. Altman, "Pharmspresso: a text mining tool for extraction of pharmacogenomic concepts and relationships from full text," *BMC bioinformatics*, vol. 10, supplement 2, p. S6, 2009.

[61] Gene Ontology Consortium, "The Gene Ontology in 2010: extensions and refinements," *Nucleic Acids Research*, vol. 38, pp. D331–D335, 2010.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 459248, 8 pages
http://dx.doi.org/10.1155/2013/459248

Research Article

On the Difference in Quality between Current Heuristic and Optimal Solutions to the Protein Structure Alignment Problem

Mauricio Arriagada,[1] and Aleksandar Poleksic[2]

[1] *Department of Computer Science, School of Engineering, Pontificia Universidad Católica de Chile, 4860 Avenue Vicuña Mackenna, 6904411 Santiago, Chile*
[2] *Department of Computer Science, University of Northern Iowa, 1227 West 27th Street, Cedar Falls, IA 50613, USA*

Correspondence should be addressed to Aleksandar Poleksic; poleksic@cs.uni.edu

Received 10 September 2012; Accepted 2 November 2012

Academic Editor: Tun-Wen Pai

The importance of pairwise protein structural comparison in biomedical research is fueling the search for algorithms capable of finding more accurate structural match of two input proteins in a timely manner. In recent years, we have witnessed rapid advances in the development of methods for approximate and optimal solutions to the protein structure matching problem. Albeit slow, these methods can be extremely useful in assessing the accuracy of more efficient, heuristic algorithms. We utilize a recently developed approximation algorithm for protein structure matching to demonstrate that a deep search of the protein superposition space leads to increased alignment accuracy with respect to many well-established measures of alignment quality. The results of our study suggest that a large and important part of the protein superposition space remains unexplored by current techniques for protein structure alignment.

1. Introduction

Pairwise protein structure alignment is one of the most important problems in computational molecular biology. At the same time, protein structure alignment is a very difficult problem, due to an infinite number of possible ways to position a pair of proteins in the three-dimensional space. Because of the enormous size of the search space, the research into protein structure alignment has been traditionally focused on the development of methods with better objective functions, that explore a relatively small but representative set of proteins' spatial superpositions.

In this paper, we take a different approach and study the benefits of searching proteins' superpositions in a more detailed manner. We demonstrate significant increase in the alignment accuracy of several well-known distance-based alignment methods, obtained by utilizing the superpositions that rigorously optimize a very simple and intuitive alignment metric, defined as the largest number of residues from the input proteins that can be fit under a predefined distance cutoff.

The size of gap between the accuracy of current heuristic solutions and optimal solutions, observed in this study, suggests that the protein structure alignment problem will likely remain a hot topic in years to come.

2. Materials and Methods

Our study is carried out using two protein structure alignment benchmarks: *Sisyphus* and *FSSP*. In both benchmarks, an in-house algorithm, MaxPairs [1], is applied to compute the superpositions that closely approximate the measure $CA \leq d$, which is defined as the largest number of pairs of residues from the input proteins that can be fit under d Ångströms. MaxPairs algorithm is based on the approximation algorithm EPSILON-OPTIMAL [1], which is capable of finding a superposition of the input proteins that fits at least as many pairs of residues under the distance $d + \varepsilon$ as an optimal superposition fits under the distance d, for any accuracy threshold $\varepsilon > 0$. As an approximation algorithm, EPSILON-OPTIMAL suffers from high computational complexity. The algorithm's run time is a high degree polynomial

in the lengths of the structures being compared. To circumvent high computational cost, the present study utilizes MaxPairs—a heuristic version of EPSILON-OPTIMAL that searches through a relatively small subset of the space of all superpositions of the input proteins inspected by EPSILON-OPTIMAL. While still not practical, as demonstrated in [1], MaxPairs enjoys accuracy superior to that of some widely utilized alignment programs and, as such, this algorithm is an indispensable tool for assessing the precision of more efficient and more popular algorithms. In present study, we set the distance cutoff to $d = 3$ Å and the accuracy threshold to $\varepsilon = 1$. Going below $\varepsilon = 1$ proves to be computationally prohibitive with our computing infrastructure.

We evaluated the performance of three well-known methods for protein structure comparison, STRUCTAL [2–4], TM-align [5], and LOCK2 [6, 7], before and after replacing their original superpositions with superpositions that optimize $CA \leq d$.

It is important to emphasize that our experiment is not designed to compare these three methods head-to-head, but rather to assess the extent of improvements in the accuracy of each method that can be made by exploring the search space in a more thorough manner.

In choosing the methods for our study, we only considered the availability of software and the simplicity of implementing the alignment scoring functions (see the Results section). An overview of the three algorithms is given below.

STRUCTAL. The STRUCTAL algorithm [2–4] employs iterative dynamic programming to balance the *cRMS* score with the lengths of aligned regions. In each iteration, the algorithm computes an optimal residue-residue correspondence (alignment) of the input proteins $a = (a_1, \ldots, a_m)$ and $b = (b_1, \ldots, b_n)$ and then finds a superposition that minimizes *cRMS* of the aligned subchains $(a_{i_1}, \ldots, a_{i_k})$ and $(b_{i_1}, \ldots, b_{i_k})$. The *cRMS* score is given by

$$cRMS = \sqrt{\frac{1}{k}\sum_{r=1}^{k}\left\|a_{i_r} - b_{i_r}\right\|^2}. \quad (1)$$

The alignment step in STRUCTAL is carried out using a dynamic programming routine, which implements the following recurrence formula:

$$D(i,j) = \max\begin{cases} D(i-1, j-1) + S(i,j) \\ D(i-1, j) - 10 \\ D(i, j-1) - 10 \end{cases} \quad (2)$$

$$D(i,0) = D(0,j) = 0,$$

where

$$S(i,j) = \frac{20}{1 + \left(d_{i,j}{}^2/5\right)}, \qquad d_{i,j} = \left\|a_i - b_j\right\|. \quad (3)$$

The outputs of STRUCTAL are the subchains p of a and q of b, along with the rigidly transformed protein b, denoted by $\hat{b}$, and a residue-residue correspondence that maximizes the STRUCTAL score

$$\sum_{i=1}^{k}\frac{20}{1 + \left\|a_{p_i} - \widehat{b_{q_i}}\right\|^2/5} - 10G_{p,q}, \quad (4)$$

where $G_{p,q}$ denotes the total number of gaps in the alignment. The STRUCTAL program used in our analysis was downloaded from http://csb.stanford.edu/levitt/Structal/.

TM-align. TM-align is another popular protein structure alignment program, widely used in many applications, in particular for assessing the quality of protein models generated by comparative modeling or abinitio techniques. The score matrix in TM-align is protein-length specific and is defined as

$$S(i,j) = \frac{1}{1 + \left(d_{i,j}/d_0\right)^2}, \quad (5)$$

where $d_0 = 1.24\sqrt[3]{L - 15} - 1.8$, and L is the length of the shorter structure [5]. In contrast to linear gap penalties employed by STRUCTAL, the gap penalties in TM-align are affine and are set to 0.6 for gap-opening and 0.0 for gap-extension [5]. An improved version of the algorithm, called Fr-TM-align, has been published [8]. The TM-align software, used in this study, was downloaded from http://zhanglab.ccmb.med.umich.edu/TM-align/.

LOCK2. LOCK2 [6] is an improved version of the original LOCK program [7]. It incorporates secondary structure information into the alignment process. An initial superposition is obtained by comparing the vectors of secondary structure elements. An iterative procedure is then applied to minimize *RMSD* between aligned subchains of the input proteins, using the threshold distance of 3 Å for atomic superposition. Rigid body motions for *RMSD* minimization are realized using quaternion transformations [9, 10].

The alignment returned by LOCK2 is a sequence of pairs of points

$$\left(a_{i_1}, b_{i_1}\right), \ldots, \left(a_{i_k}, b_{i_k}\right), \quad i_1 < \cdots < i_k, \quad (6)$$

where a_{i_r} are b_{i_r} each other's nearest neighbors. More specifically, for every $r \in \{1, \ldots, k\}$, the point b_{i_r} is the closest point in protein b to the point a_{i_r} and vice versa. The final alignment is generated through a two-step process. First, for every atom a_i from protein a, the algorithm finds the nearest atom from protein b that is at distance ≤3 Å from a_i. In the second step, the algorithm selects the maximum number of aligned pairs in sequential order, by removing pairs that violate colinearity.

The LOCK2 software can be downloaded from http://lock2.stanford.edu.

3. Results

3.1. Sisyphus Benchmark. The *Sisyphus* test [11] is frequently used to assess the accuracy of automated methods for

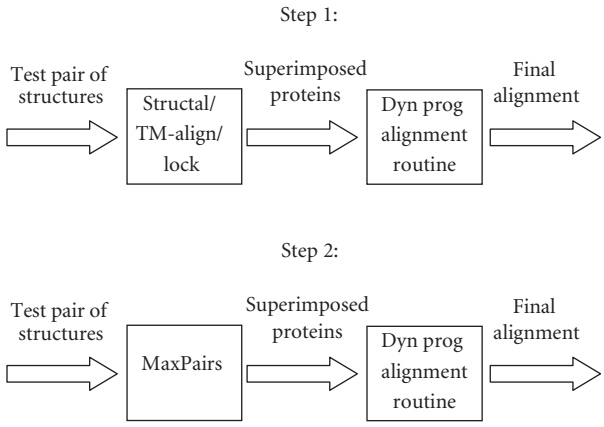


Figure 1: The procedure for creating methods' specific alignments and alignments based on MaxPairs superpositions.

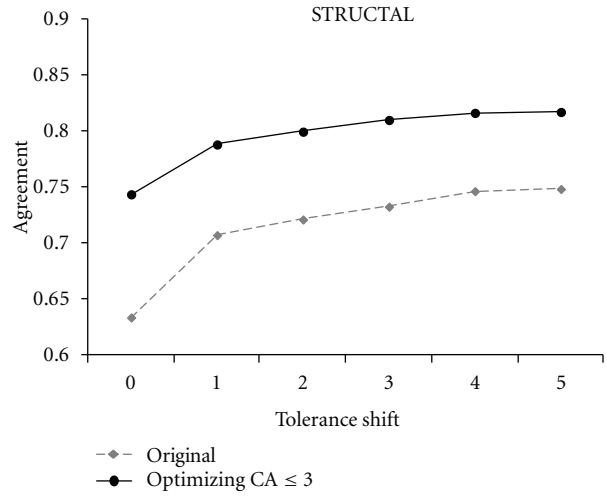


Figure 2: The accuracy of the STRUCTAL algorithm using original versus optimized superpositions in the *Sisyphus* benchmark.

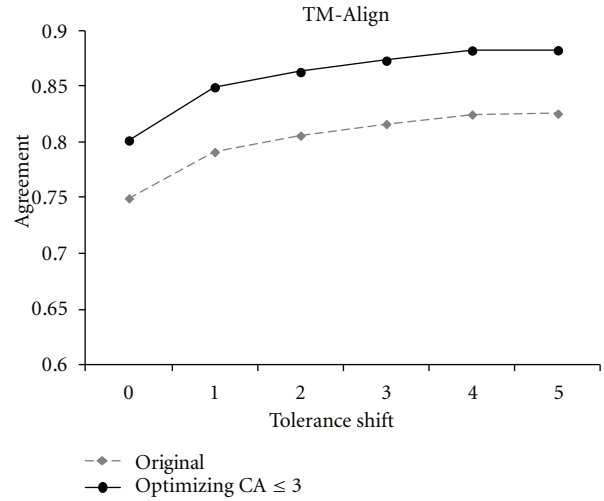


Figure 3: The agreement of TM-align alignments and reference alignments in the *Sisyphus* benchmark.

protein structure comparison [1, 12]. This sophisticated benchmark utilizes 125 alignments of structurally related proteins, created by experts in the field of protein structure analysis. The reference alignments can be downloaded from http://sisyphus.mrc-cpe.cam.ac.uk.

In present study, we (like Rocha et al. [12]) utilize only a subset of the *Sisyphus* test set, containing 106 alignments between single-chain proteins. The two-step process is illustrated in Figure 1. In the first step, STRUCTAL, TM-align, and LOCK2 are run with default parameters to generate the methods' specific alignments between proteins from the *Sisyphus* set. These alignments are then compared to the reference ("gold-standard") alignments to compute the percentage of correctly aligned residue pairs [1, 12].

In the second step, the MaxPairs algorithm is run to compute the set of (near-)optimal superpositions, namely, the superpositions that rigorously maximize the number of pairs of atoms that can be fit under 3 Å. We used our own implementations of the STRUCTAL, TM-align, and LOCK2 alignment procedures to compute optimal residue-residue correspondence (alignment) between the newly superimposed proteins. The percentage agreement with reference alignments is recorded again and compared to the agreement obtained in the first step.

The agreement with reference alignments in the *Sisyphus* test is defined as a function of the magnitude of the alignment error. More specifically, for the alignment tolerance shift s, the agreement is defined as I_s/L_{ref}, where I_s is the number of aligned residues that are shifted by no more than s positions in the reference alignment and L_{ref} is the length of the reference alignment [12]. The perfect agreement is the one that corresponds to zero-shift ($s = 0$).

The dashed lines in Figures 2, 3, and 4 track the performance of original STRUCTAL, TM-align, and LOCK2 methods. The solid lines show the performance of the same methods when run on the superpositions that maximize the number of residues under 3 Å. As seen in these figures, there is a significant boost in the methods' accuracy resulting from the "fine-tooth comb" search of superposition space. More precisely, the new superpositions improve absolute agreement with the reference alignments for STRUCTAL, TM-align, and LOCK2 by 11%, 5%, and 5%, respectively, with a similar trend continuing for nonzero shift.

The increase in number of correctly aligned residues, obtained by switching to MaxPairs superpositions, varies from one pair of structures to another (Figures 5, 6, and 7). For some pairs, the difference is striking. However, it should be emphasized that, in some of these cases, such a high difference might be due to unavailability of information in PDB files used by the methods in our study. For instance, the LOCK method is built to take advantage of the residues' secondary structure assignment. Hence, it is reasonable to assume that the lack of secondary structure information in the PDB file for one or both structures will often decrease the accuracy of the LOCK alignment of those structures.

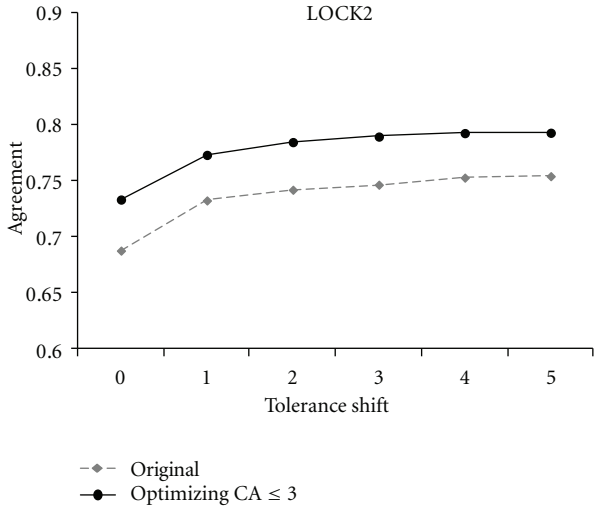


Figure 4: The agreement of LOCK2 alignments and reference alignments in the *Sisyphus* benchmark.

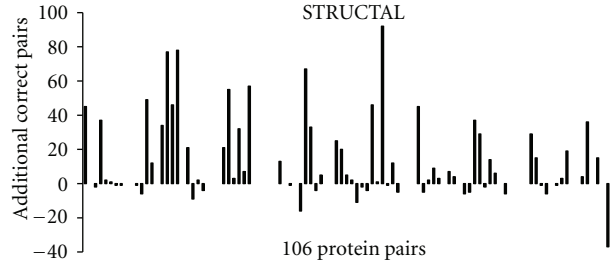


Figure 5: The increase in accuracy of STRUCTAL obtained on 106 pairs from the *Sisyphus* benchmark.

A more detailed analysis shows that, when MaxPairs superpositions are used, the number of residue pairs correctly aligned by STRUCTAL increases by more than 10 for 31 out of 106 test pairs. The corresponding number of test pairs for which the same magnitude of increase is observed for TM-align and LOCK is 14 and 13, respectively. For comparison, original STRUCTAL superpositions have such an advantage only in 3 out of 106 test pairs. For TM-align and LOCK, the corresponding numbers are 5 and 4.

The value added by the deep search of superposition space makes some of the methods analyzed here comparable to the best to date methods evaluated in the *Sisyphus* test. A slight accuracy advantage of algorithms such as Matt [13], PPM [14], and ProtDeform [12] is due to the fact that these methods consider proteins as flexible, rather than rigid objects. In other words, unlike STRUCTAL, TM-align, and LOCK2, which all utilize single transformations of input proteins to compute final alignments, the new generation of protein structure alignment methods consider sequences of different rigid transformations at different sites. It should be emphasized that the methods based on sequences of local transformations can themselves benefit from incorporating the "fine-tooth comb" search to detect fragments of local

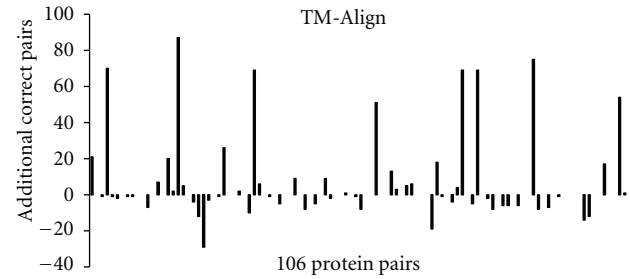


Figure 6: The increase in accuracy of TM-align obtained on 106 pairs from the *Sisyphus* benchmark.

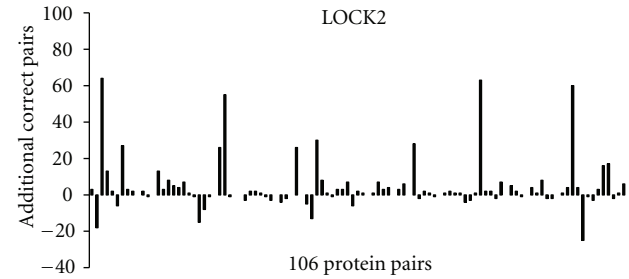


Figure 7: The increase in accuracy of LOCK2 obtained on 106 pairs from the *Sisyphus* benchmark.

similarity. This would lead to further improvements in their overall accuracy, but the true extent of these improvements can only be accessed through a carefully designed study.

3.2. FSSP Benchmark. Our second benchmarking set utilizes 183 representative pairs of proteins, related at various levels according to FSSP structural classification [15]. This test set consists of 55 family pairs, 68 superfamily pairs, and 60-fold pairs (see Supplementary Material available online at doi:10.1155/2012/459248).

In contrast to *Sisyphus* benchmark, which compares alignments returned by automated methods to those generated by human experts, the alignment precision in the FSSP benchmark is assessed using a set of well-known alignment quality measures:

(i) NumPairs(d) represents the number of aligned pairs of residues in two proteins that are at distance $\leq d$ Ångströms from each other. We note that, unlike $CA \leq d$, which is a globally optimal metric, representing the maximum number of pairs of residues in the superimposed structures that can be placed under d Ångströms, NumPairs(d) represents the method specific count of pairs of aligned residues at distance $\leq d$.

(ii) *Similarity Index*, denoted by SI, is defined as $cRMS \times \min\{L(a), L(b)\}/Nmat$, where $Nmat$ is the number of aligned residues in proteins a and b and $L(a)$ and $L(b)$ are the lengths of a and b, respectively [16]. The $cRMS$ score, used in the formula for SI, is computed based upon the method specific alignments.

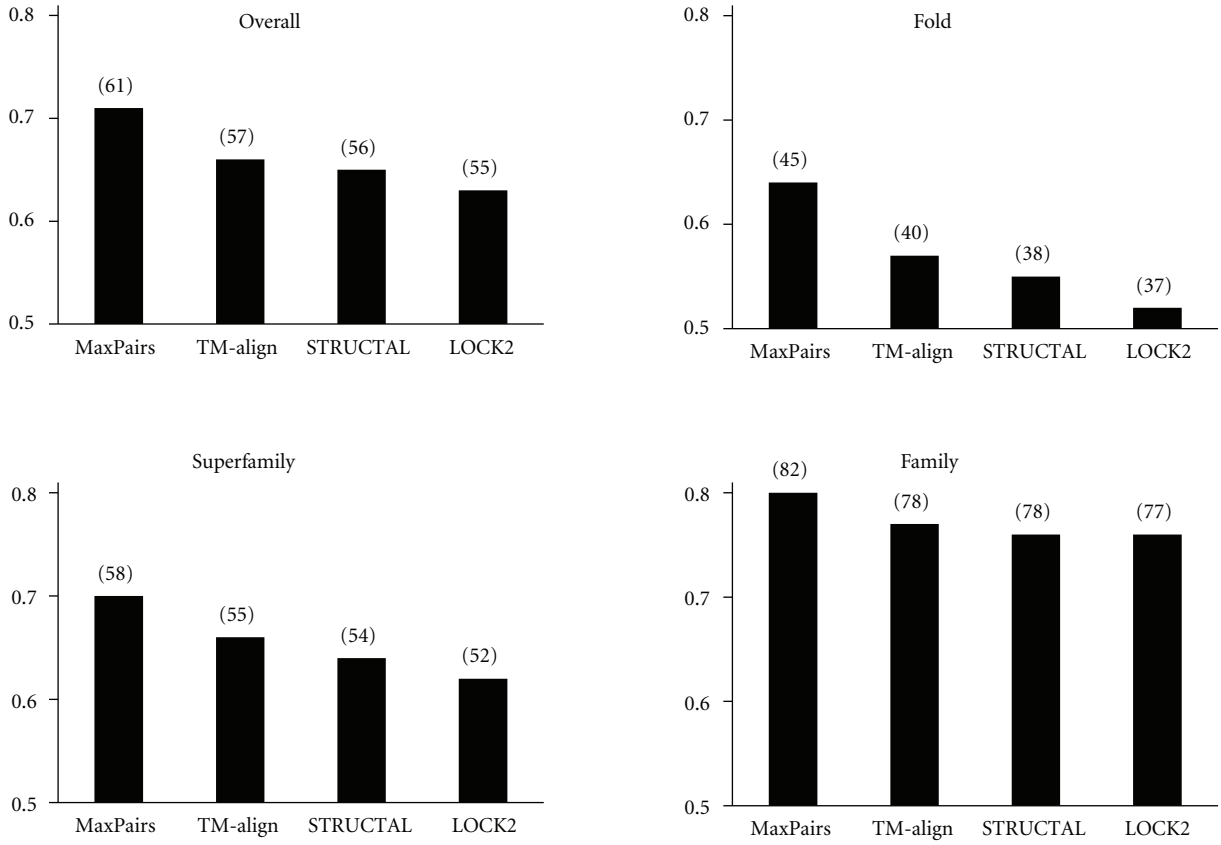


Figure 8: Alignment independent PSI scores in the FSSP benchmark. The number in parentheses is the highest number of pairs of residues (averaged over each test set) that can be placed under 3 Å, given the superpositions generated by each method.

Table 1: Average (per-pair) accuracy of STRUCTAL, TM-align, and LOCK2 in the FSSP benchmark, for all structural levels combined. The best results are indicated in bold.

	NumPairs(3)	PSI(3)	SI
STRUCTAL			
Original	50.47	0.59	7.85
Near-optimal	**53.81**	**0.63**	**7.37**
TM-align			
Original	53.35	0.62	**5.86**
Near-optimal	**55.85**	**0.65**	5.95
LOCK2			
Original	51.75	0.60	8.35
Near-optimal	**58.46**	**0.68**	**5.69**

(iii) *The Percentage of Structural Similarity*, PSI(d), is defined as NumPairs(d)/ min$\{L(a), L(b)\}$ (see, for example, [8]).

As seen in Table 1, a more detailed search of the superposition space increases both NumPairs and PSI scores for all three methods in our study. The increase in SI scores is also seen for both STRUCTAL and LOCK2. It is interesting to note, though, that the original TM-align superpositions yeald better SI scores than the optimal superpositions.

Table 2: Average accuracy of the three methods in our study, computed on 60 pairs of proteins that share the same FSSP fold.

	NumPairs(3)	PSI(3)	SI
STRUCTAL			
Original	31.48	0.47	10.07
Near-optimal	36.68	0.54	9.63
TM-align			
Original	35.98	0.52	7.60
Near-optimal	39.58	0.57	7.76
LOCK2			
Original	34.82	0.50	12.56
Near-optimal	42.47	0.61	7.25

The FSSP level-specific results of our benchmarking analysis are summarized in Tables 2, 3, and 4.

Figure 8 shows the alignment independent PSI scores computed from superpositions generated by STRUCTAL, TM-align, and LOCK2. For reference, a near-optimal PSI score, averaged across the FSSP test set and computed by the MaxPairs algorithm, is also provided in this figure.

The data used in Figure 8 shows that (on average) STRUCTAL, TM-align, and LOCK fail to place 8%, 7%, and 11% pairs of residues at distance $\leq$ 3 Å, respectively.

Table 3: Average accuracy computed on the set of 68 pairs of proteins that belong to the same FSSP superfamily.

	NumPairs(3)	PSI(3)	SI
STRUCTAL			
Original	47.71	0.58	8.41
Near-optimal	50.09	0.61	7.61
TM-align			
Original	51.40	0.62	6.11
Near-optimal	52.71	0.64	6.10
LOCK2			
Original	48.63	0.59	8.17
Near-optimal	55.37	0.67	5.85

Table 4: Average accuracy computed on the set of 55 pairs of structures from the same FSSP family.

	NumPairs(3)	PSI(3)	SI
STRUCTAL			
Original	74.6	0.74	4.76
Near-optimal	77.11	0.76	4.62
TM-align			
Original	74.71	0.73	3.65
Near-optimal	77.49	0.76	3.79
LOCK2			
Original	74.09	0.73	3.98
Near-optimal	79.73	0.78	3.80

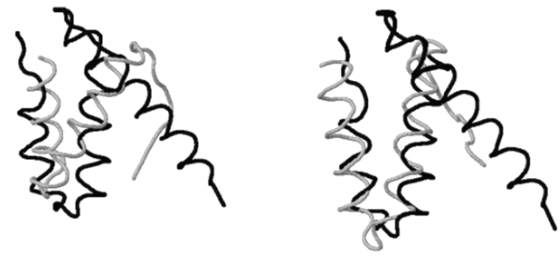
As expected, the best performance of these methods is observed at the FSSP family level (STRUCTAL fails to place 5%, TM-align: 5%, LOCK: 6%) and worst at FSSP fold level (STRUCTAL: 15%, TM-align: 12%, LOCK: 17%).

3.3. Illustrative Examples. Several examples illustrating the advantage of the deep search of superposition space are given in Figures 9, 10, 11, 12, and 13.

While examples in Figures 9–13 are striking, it should be noted that they represent rather isolated cases. In fact (as the reader can conclude from Figures 5, 6, and 7), there are several examples where the output of heuristic methods compares favorably to that of MaxPairs (although the difference in quality is not as obvious as that shown in Figures 9–13). As emphasized before, in many instances, the inaccuracy of the alignment generated by heuristic methods is due to insufficient structural information stored in the PDB file, relied upon these methods.

(a) (b)

Figure 9: Structural alignment of two cystatin-like folds: *delta-5-3-ketosteroid isomerase from pseudomonas putida*, PDB ID: 1opy (black) and *chicken egg white cystatin*, PDB ID: 1cewI (gray), obtained by (a) a heuristic method and (b) MaxPairs. For simplicity of presentation, we hide the parts of two structures that are not well superimposed by either program. In this particular test case, switching to MaxPairs superpositions yields a twofold increase in the number of pairs of residues that can be fit under 3 Å (59 versus 29). The corresponding percentage increase for the distance cutoff of 5 Å is 76% (65 versus 37 pairs).

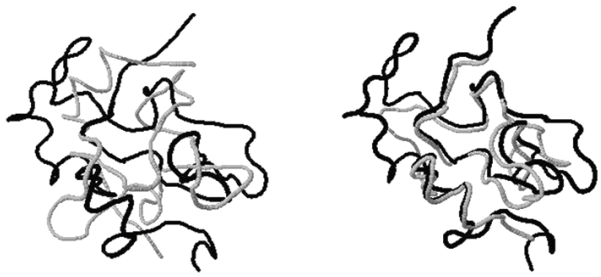


(a) (b)

Figure 10: Structural superposition of the *urease from Bacillus pasteurii*, PDB ID: 1ubpA (black) and the *dynein light chain 1 from human*, PDB ID: 1cmiA (gray), obtained by (a) a heuristic method and (b) MaxPairs (we hide misaligned C-terminal regions from both structures). In this test case, a subtle change in structural superposition, made by MaxPairs, increases the number of pairs of residues that can be fit under 3 Å from 9 to 36 (and from 16 to 42 when the cutoff distance of 5 Å is used).

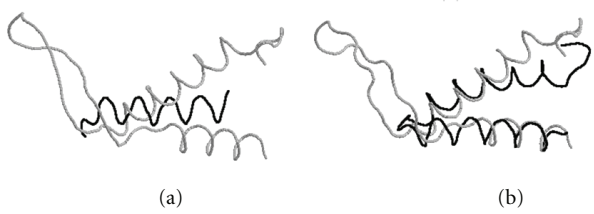
4. Discussion

Resent years have witnessed advances in the development of methods for approximate and exact solution to protein structure alignment problem. One of the first such methods is the Umeyama's algorithm for finding the transformation that gives the least mean squared error between two point patterns [17]. Since then, several algorithms have been published for finding a near-optimal solution to the structure alignment problem under distance constraints. The procedure by Akutsu, for example, returns a superposition of the input proteins that fits at least as many pairs of residues under the distance $c \cdot d$ as an optimal alignment fits under the distance d, for every fixed $c > 1$ [18]. This algorithm runs on the order of $O(n^8)$, where n denotes the protein length. An improved running time procedure for the same problem has also been published [19]. The EPSILON-OPTIMAL algorithm, used in present study, is able to place at least as many pairs of residues under the distance $d + \varepsilon$ as an optimal superposition places under the distance d. The asymptotic cost of EPSILON-OPTIMAL is $O(n^4)$ for globular and $O(n^8)$ for nonglobular proteins [1].



(a) (b)

Figure 11: Structural superposition of *HiPIP* (high-potential iron protein) from Chromatium *vinosum*, PDB ID: 1ckuA (black) and HiPIP isolated from the phototrophic bacterium Rhodocyclus *tenuis*, PDB ID: 1isuA (gray), obtained by (a) a heuristic method and (b) MaxPairs. This is another example illustrating an obvious difference in quality of two structural matches.

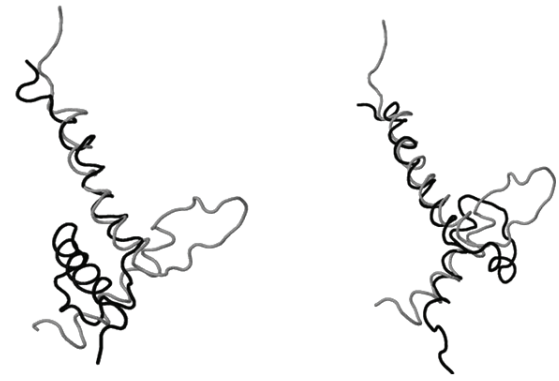


(a) (b)

Figure 12: Structural superposition of two helical regions in the *elongation factor TS*, PDB ID: 1tfeA (black) and the *ribosomal protein S7*, PDB ID: 1rssA (gray), obtained by (a) a heuristic method and (b) MaxPairs. Regions not aligned well by the two programs are hidden for simplicity of presentation. When run on superpositions generated by MaxPairs, the same heuristic method aligns 22 more residues under the distance 3 Å (29 versus 7).

The polynomial time approximation schemes (PTASs) have been designed for selected nonsequential protein structure alignment measures [20] as well as for the class of measures satisfying the so-called Lipschitz condition [21]. Moreover, methods exist that rigorously minimize proteins' intra-atomic distances, including the algorithm by Caprara et al., which is capable of approximating the "Contact Map Overlap" (CMO) measure with great accuracy [22]. Finally, the algorithms for absolute optimum, with respect to selected alignment metrics, have also been published [1, 23], but they are computationally too expensive for everyday use.

Although inefficient for large scale analysis, the algorithms for exact solution are indispensable tools for assessing the accuracy of more commonly used heuristic methods. The present study utilizes a set of precomputed superpositions to evaluate the improvements in accuracy of three well-known protein structure alignment algorithms, obtained by the deep search of the superposition space. In the *Sisyphus* benchmark, these superpositions increase the accuracy of alignments generated by STRUCTAL, TM-align, and LOCK2



(a) (b)

Figure 13: Structural superposition of the *DNA-binding domain of PHO4*, PDB ID: 1a0aA (black) and the *basic/helix-loop-helix/leucine* zipper domain of the upstream stimulatory factor, PDB ID: 1an4A (gray), obtained by (a) a heuristic method and (b) MaxPairs. Unlike the heuristic method, MaxPair is capable of aligning both helical regions from these two proteins.

by 11%, 7%, and 6%, respectively. An improvement of similar magnitude is seen after allowing for alignment errors (residue shifts). In the FSSP benchmark, the new superpositions increase NumPairs and PSI scores for STRUCTAL, TM-align, and LOCK2 by ~7%, ~5%, and ~13%, respectively. A particularly noticeable improvement is seen in the *Similarity Index* scores of alignments generated by LOCK2 (from 8.35 to 5.69).

We emphasize that our analysis provides an estimate of the lower bound on the difference between optimal and heuristic solution, since alignments generated by MaxPairs are not always optimal (in the strict sense).

Finally, it is reasonable to expect that a more thorough exploration of the superposition space, coupled with the fragment-based alignment techniques, can be used to further improve the precision of methods based on sequences of local transformations, such as Matt [13], PPM [14], and ProtDeform [12].

5. Conclusions

A typical distance-based protein structure alignment method explores the space of proteins' spatial superpositions, computing an optimal residue-residue correspondence (alignment) each time a new superposition is generated. Because of the large search space, current methods for protein structure alignment must trade precision for speed and explore only a small but representative set of superpositions.

We utilize an algorithm capable of finding an alignment of any specified accuracy to demonstrate significant increase in the alignment quality of solutions generated by three popular protein structure alignment methods, obtained through the deep search of the superposition space. The large lower bound on the size of gap between optimal and heuristic solutions,

observed in this study, suggests that the protein structure alignment problem will likely remain an attractive research area throughout the next decade.

Acknowledgment

A. Poleksic was supported, in part, by a Professional Development Assignment from the University of Northern Iowa.

References

[1] A. Poleksic, "Algorithms for optimal protein structure alignment," *Bioinformatics*, vol. 25, no. 21, pp. 2751–2756, 2009.

[2] S. Subbiah, D. V. Laurents, and M. Levitt, "Structural similarity of DNA-binding domains of bacteriophage repressors and the globin core," *Current Biology*, vol. 3, no. 3, pp. 141–148, 1993.

[3] M. Gerstein and M. Levitt, "Using iterative dynamic programming to obtain accurate pairwise and multiple alignments of protein structures," in *Proceedings of the 4th International Conference on Intelligent Systems for Molecular Biology*, pp. 59–67, AAAI Press, Menlo Park, Calif, USA, 1996.

[4] M. Levitt and M. Gerstein, "A unified statistical framework for sequence comparison and structure comparison," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 11, pp. 5913–5920, 1998.

[5] Y. Zhang and J. Skolnick, "TM-align: a protein structure alignment algorithm based on the TM-score," *Nucleic Acids Research*, vol. 33, no. 7, pp. 2302–2309, 2005.

[6] J. Shapiro and D. Brutlag, "FoldMiner: structural motif discovery using an improved superposition algorithm," *Protein Science*, vol. 13, no. 1, pp. 278–294, 2004.

[7] A. P. Singh and D. L. Brutlag, "Hierarchical protein structure superposition using both secondary structure and atomic representations," in *Proceedings of the International Conference on Intelligent Systems for Molecular Biology*, vol. 5, pp. 284–293, 1997.

[8] S. B. Pandit and J. Skolnick, "Fr-TM-align: a new protein structural alignment method based on fragment alignments and the TM-score," *BMC Bioinformatics*, vol. 9, article 531, 2008.

[9] B. K. P. Horn, "Closed-form solution of absolute orientation using unit quaternions," *Journal of the Optical Society of America*, vol. 4, pp. 629–642, 1997.

[10] B. K. P. Horn and H. M. Hilden, "Closed-form solution of absolute orientation using orthonormal matrices," *Journal of the Optical Society of America*, vol. 5, pp. 1127–1135, 1998.

[11] A. Andreeva, A. Prlić, T. J. P. Hubbard, and A. G. Murzin, "SISYPHUS—structural alignments for proteins with non-trivial relationships," *Nucleic Acids Research*, vol. 35, no. 1, pp. D253–D259, 2007.

[12] J. Rocha, J. Segura, R. C. Wilson, and S. Dasgupta, "Flexible structural protein alignment by a sequence of local transformations," *Bioinformatics*, vol. 25, no. 13, pp. 1625–1631, 2009.

[13] M. Menke, B. Berger, and L. Cowen, "Matt: local flexibility aids protein multiple structure alignment," *PLoS Computational Biology*, vol. 4, no. 1, article e10, 2008.

[14] G. Csaba, F. Birzele, and R. Zimmer, "Protein structure alignment considering phenotypic plasticity," *Bioinformatics*, vol. 24, no. 16, pp. i98–i104, 2008.

[15] L. Holm, C. Ouzounis, C. Sander, G. Tuparev, and G. Vriend, "A database of protein structure families with common folding motifs," *Protein Science*, vol. 1, no. 12, pp. 1691–1698, 1992.

[16] R. Kolodny, P. Koehl, and M. Levitt, "Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures," *Journal of Molecular Biology*, vol. 346, no. 4, pp. 1173–1188, 2005.

[17] S. Umeyama, "Least-squares estimation of transformation parameters between two point patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 4, pp. 376–380, 1991.

[18] T. Akutsu, "Protein structure alignment using dynamic programming and iterative improvement," *IEICE Transactions on Information and Systems*, vol. E79-D, no. 12, pp. 1629–1636, 1996.

[19] S. C. Li and Y. K. Ng, "On protein structure alignment under distance constraint," in *Proceedings of ISAAC*, pp. 65–76, 2009.

[20] J. Xu, F. Jiao, and B. Berger, "A parameterized algorithm for protein structure alignment," *Journal of Computational Biology*, vol. 14, no. 5, pp. 564–577, 2007.

[21] R. Kolodny and N. Linial, "Approximate protein structural alignment in polynomial time," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 33, pp. 12201–12206, 2004.

[22] A. Caprara, R. Carr, S. Istrail, G. Lancia, and B. Walenz, "1001 Optimal PDB structure alignments: integer programming methods for finding the maximum contact map overlap," *Journal of Computational Biology*, vol. 11, no. 1, pp. 27–52, 2004.

[23] C. Ambühl, S. Chakraborty, and B. Gärtner, "Computing largest common point sets under approximate congruence," in *Proceedings of the ESA*, vol. 1879 of *Lecture Notes in Computer Science*, pp. 52–64, 2000.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 106407, 6 pages
http://dx.doi.org/10.1155/2013/106407

Review Article

Cancer Vaccines: State of the Art of the Computational Modeling Approaches

Francesco Pappalardo,[1] Ferdinando Chiacchio,[2] and Santo Motta[3]

[1] *Dipartimento di Scienze del Farmaco, Università degli Studi di Catania, V.le A. Doria 6, 95125 Catania, Italy*
[2] *Dipartimento di Ingegneria Elettrica Elettronica e Informatica, Università degli Studi di Catania, V.le A. Doria 6, 95125 Catania, Italy*
[3] *Dipartimento di Matematica e Informatica, Università degli Studi di Catania, V.le A. Doria 6, 95125 Catania, Italy*

Correspondence should be addressed to Francesco Pappalardo; francesco@dmi.unict.it

Received 19 October 2012; Accepted 20 November 2012

Academic Editor: Hao-Teng Chang

Cancer vaccines are a real application of the extensive knowledge of immunology to the field of oncology. Tumors are dynamic complex systems in which several entities, events, and conditions interact among them resulting in growth, invasion, and metastases. The immune system includes many cells and molecules that cooperatively act to protect the host organism from foreign agents. Interactions between the immune system and the tumor mass include a huge number of biological factors. Testing of some cancer vaccine features, such as the best conditions for vaccine administration or the identification of candidate antigenic stimuli, can be very difficult or even impossible only through experiments with biological models simply because a high number of variables need to be considered at the same time. This is where computational models, and, to this extent, immunoinformatics, can prove handy as they have shown to be able to reproduce enough biological complexity to be of use in suggesting new experiments. Indeed, computational models can be used in addition to biological models. We now experience that biologists and medical doctors are progressively convinced that modeling can be of great help in understanding experimental results and planning new experiments. This will boost this research in the future.

1. Introduction

Vaccines for cancer represent an alternative approach to the use of standard drugs. Differently from the traditional vaccines that prevent disease instructing the immune system on how to recognize and destroy a particular pathogen, cancer vaccines enlist the patient's immune system to destroy existing cancer cells. While simple in concept, the development of products has proven difficult. Problems specifically lie in eliciting sufficient, tumor-selective stimulation of an immune system that is already tolerant of cancer cells [1, 2].

Revolutions in biotechnology and information technology have produced enormous amounts of data and are accelerating the extension of our knowledge of biological systems. These advances are changing the way biomedical research, development, and applications are done. Clinical data complement biological data, enabling detailed descriptions of various healthy and diseased states, progression, and responses to therapies. The availability of data representing various biological states, processes, and their time dependencies enable the study of biological systems at various levels of organization, from molecule to organism, and even population levels.

Specific systems biology models, that is, applications of computer and mathematical models that enable the simulation of biological processes, can be used to investigate the physiology and pathology of the immune responses involved in vaccination and immunotherapy. This involves applications of computational simulations to the discovery, design, and optimization of vaccines and other immunotherapies.

The term immunotherapies usually refers to the treatment of established disease while the term vaccine is restricted to prophylactic immune interventions. We will use "vaccines" to refer to generic immune intervention and will use terms "therapeutic vaccines" and "prophylactic vaccines"

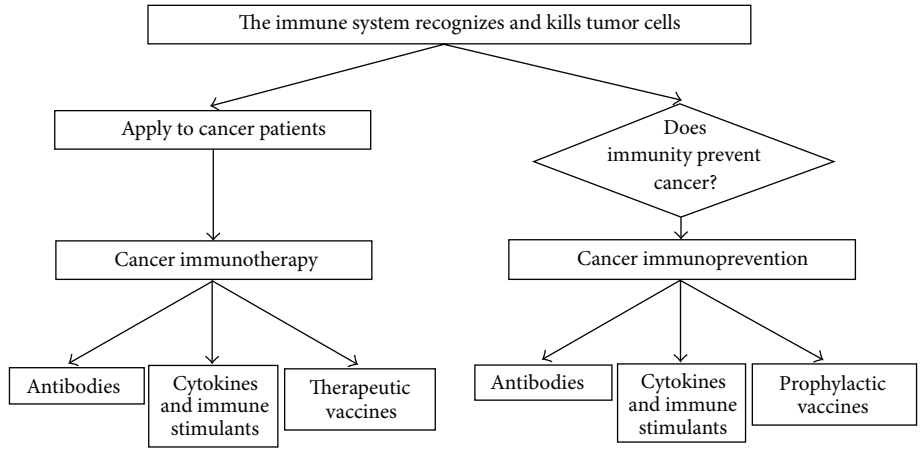


Figure 1: Tumor immunology and the main difference between cancer immunotherapy and cancer immunoprevention.

to distinguish between the two modalities. Vaccine design is amenable to the application of modeling techniques, for both the discovery and development of new and existing vaccines. In what follows, we first deal with a brief overview of different types of existing cancer vaccines; we then focus on modeling cancer vaccines and finally we draw our final remarks.

2. A Brief Overview of Cancer Vaccines

The ultimate aim of a vaccine is to activate a component of the immune system such as B lymphocytes, which produce antibodies or T lymphocytes, which directly kill tumor cells. Antibodies must recognize antigens in the native protein state on the cell's surface. Once bound, antibodies are capable of destroying tumor cells by means of antibody-dependent cellular cytotoxicity or complement-mediated cytotoxicity. T lymphocytes recognize proteins as major histocompatibility complex complexed with peptides that can vary in size, presented on the surface of the cells recognized.

Recent research [3] demonstrated that the vaccine approach may also be useful in the prevention and treatment of cancer (tumor immunology, see Figure 1). It is known that the immune system eliminates most of the cancer cells (cancer immunoediting [4]). Those that are not recognized escape immune surveillance, leading to tumors. Tumor vaccines can thus be used to stimulate an immune response against poorly immunogenic tumor variants. In few words, the ultimate goal of tumor immunology is to understand the interactions between tumor and immune system cells and to devise immune based approaches to fight cancer.

The use of cytotoxic T cells (CTLs), dendritic cells (DC), and antibodies, actually represent well-known approaches in cancer immunotherapy [5].

The use of anti-idiotype (Id) antibodies as vaccines to stimulate immune system response against tumors, have been demonstrated effective in preventing tumor growth and curing mice with established tumors [6]. Several monoclonal anti-Id antibodies that have the appearance of distinct human tumor-associated antigens (TAAs) have been developed and tested in the clinic, demonstrating good results. Indeed the efficacy of these vaccines will depend on the results of several Phase III clinical trials. Numerous studies in mouse tumor models have shown that DCs pulsed with tumor antigens can induce protective and therapeutic anti-tumor immunity [7]. It is, however, worth to mention that the complexity of the DC system requires rational manipulation of DCs to achieve protective or therapeutic immunity.

Recently it has been shown that prophylactic vaccines administered to transgenic mice prone to cancer development can completely prevent tumor onset and restore a normal life expectancy [8]. Even though prophylactic cancer vaccines are still far from human application, this opens up an entirely new perspective in cancer prevention, leading to a future in which vaccines will equally contribute to the prevention of infectious diseases and cancer.

3. Modeling Cancer Vaccines

Computational models have been recognized as relevant for the understanding of biological systems. In particular, models are suitable for guiding biology from a qualitative to a quantitative, thus predictive, science. Pharmaceutical companies are starting to use models to optimize/predict therapeutic effects at the organism level, suggesting that computational biology can effectively play a key role in this field [9].

Obviously to model the behavior of a cancer vaccine, one needs to model the immune system that is one of the most exciting challenges as it represents one of the most complex biological systems. It is, in fact, an adaptive learning system that operates at multiple levels (molecules, cells, organs, organisms, and groups of organisms). Immunological research, both basic and applied, needs to deal with this complexity.

Computational immunologists increasingly use mathematical modeling and computer simulation to study the immune system and the immune responses to different pathogens [10]. Thus, quantitative models that appropriately

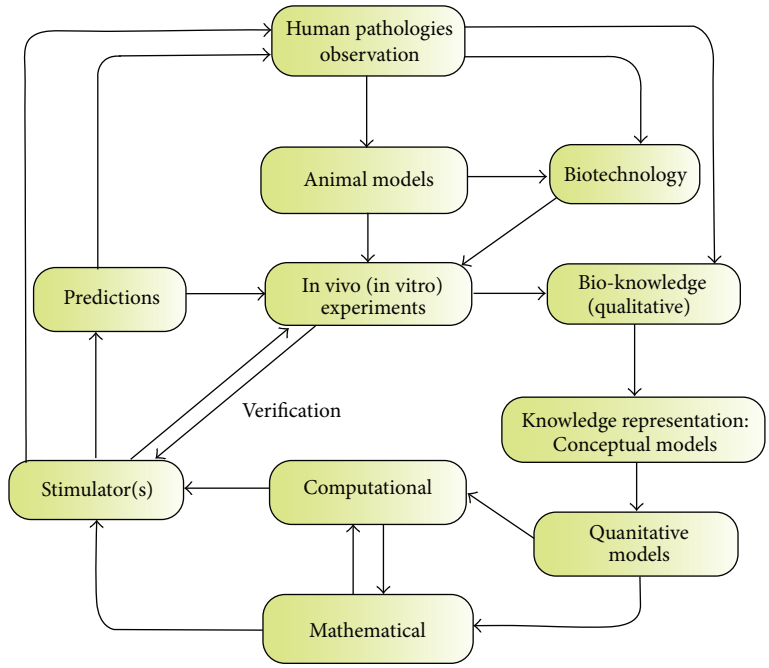


Figure 2: The modeling cycle.

capture the complexity (both in the architecture and the function) of the immune system are an integral component of the personalized medicine efforts. In silico models of the immune system can provide answers to a variety of questions, including understanding the general behavior of the immune system, the course of disease, the effects of treatments, the analysis of cellular and molecular interactions, and eventually the simulation of laboratory experiments. Here we will focus on modeling the immune response against tumors elicited by a cancer vaccine.

Figure 2 summarizes the modeling cycle that all the modeling approaches should follow.

4. The SimTriplex Model

One of the first example in modeling a cancer vaccine is represented by the SimTriplex model. It is an agent-based model specifically tailored to simulate the effects of "Triplex" tumor-preventive cell vaccines in HER-2/neu transgenic mice prone to the development of mammary carcinoma [11, 12]. The Triplex vaccine blocks mammary carcinogenesis when administered to BALB-neuT mice starting at 6 weeks of age, allowing very long (>1 y) tumor-free survival [13]. The major limitation of the very effective Triplex vaccine was that only a Chronic protocol, with more than 60 vaccinations distributed throughout the life of the mouse, blocked mammary carcinogenesis, whereas shorter and/or delayed protocols left mice exposed to tumor onset.

SimTriplex includes a variety of cellular and molecular entities, including tumor and vaccine cells, B and plasma cells, helper and cytotoxic T cells, macrophages, dendritic cells, antigens, antibodies, and cytokines. The attributes of each cell entity include position, age, and state (e.g., resting, activated, memory, antigen-presenting, etc.). Changes of state (e.g., cell activation, cytotoxicity, cell death, etc.) are governed by a set of rules based on tumor immunology.

Antigen-specific immune interactions (antibody or immunoglobulin/antigen (IG/Ag) and T cell receptor/peptide/MHC (TCR-pMHC)) are modeled with bit-strings (sequences of 0 s and 1 s). Hamming distances is used as a measure of affinity among receptors and co-receptors: the probability of an interaction depends on the number of matches.

The simulation space is a two-dimensional triangular lattice (six neighbor sites) with periodic boundary conditions. Cells and molecules are free to move across the lattice sites. At each time step, representing 8 hours of real time, cells and molecules residing on the same lattice site can interact.

To model the continuous carcinogenic process of HER-2/neu transgenic mice, new tumor cells appear in the lattice, and existing tumor cells replicate (and rarely die). The simulation stops if the total number of tumor cells exceeds a threshold, signifying the formation of a palpable tumor mass, or after a defined number of time steps, typically more than 1 year of real time.

Probabilistic elements affect various starting variables (e.g., initial positions in the lattice) and interactions (e.g., cytotoxic death of tumor cells). The outcome of each run of the simulator, entailing the generation of a large number of pseudo-casual numbers, is taken to simulate the results of one mouse, thus reproducing experimental variability between individual mice.

SimTriplex model coupled with optimization techniques (based on combinatorial optimization algorithms as genetic algorithms and simulated annealing) allowed to search for

an optimal vaccination schedule to obtain the same efficacy of the Chronic protocol with a definitively reduced vaccine administrations. Simtriplex predictions has been verified in a in-vivo experiment (probably the first model results verified in vivo). Results show that in-silico predicted schedule does significantly reduce the tumors multiplicity on the ten mice mammary glands even if the vaccination efficacy for the first appearing of tumor was still overestimated. Further adjustment of the model is required to include evidence of immune aging which appeared from in vivo follow up results [13, 14].

5. The MetastaSim Model

The Triplex vaccine proved to be effective also as a therapeutic vaccine, showing its ability to be used against induced lung metastases [15]. Briefly, lung metastases were induced in BALB-neuT mice by intravenous injection of syngeneic mammary carcinoma cells.

The administration of the vaccine started one day after the intravenous injection of the metastatic cells and it is repeated twice weekly up to the end of the experiment (day 32), with lower but good prevention rates when the same cycle is started 7 days after the induction of the metastases (Triplex+7 protocol). The immunological responses in the immunoprevention and therapeutic experiments overlap only partially. A major goal of biologists is to better understand the biological behavior to improve the efficacy of the therapeutic treatment and to try to predict, for example, the outcomes of longer experiments in order to move faster towards clinical phase I trials. In a recent work [16], we developed a new computational model named MetastaSim to be used as an in silico virtual lab can help answering these questions.

The MetastaSim model has been inspired by the SimTriplex model. MetastaSim has in common with SimTriplex the same modeling framework [17, 18] and some of the biological mechanisms shared by the in vivo experiments they model. However it has some important differences, that is, a complete revision of the cancer growth kinetics. The model is now able to simulate multiple different metastatic nodules, each one with its own growth rate, more accurately. To reproduce the growth in time of nodules, the Gompertz growth law is now used in its differential form.

An exhaustive search for any optimal protocol has been performed. Results showed that it is possible to obtain in silico a reduction of approximately 45% in the number of vaccinations. Most of the protocols presented there share a similar vaccination strategy that is composed by a boost of three vaccine injections, a period of rest, and then a series of vaccine recalls that are somewhat equally spaced. The model suggests that any optimal protocol for preventing lung metastases formation should be therefore composed by an initial massive vaccine dosage followed by few vaccine recalls. Even if this is a well-known vaccination strategy in immunology, since it is commonly used for many infectious diseases such as tetanus and hepatitisB, it can be still considered a relevant result in the field of cancer-vaccines immunotherapy.

6. Model of Immunotherapy and Cancer Vaccination

Unfortunately, the efficacy of available therapeutic strategies for cancer still remain poor. Moreover, widely adopted approaches to cure or, at least, delay cancer development that is, chemotherapy and radiotherapy, both still carry major side effects for individual patients. In order to better understand therapies, experimentalists and clinicians are increasingly appreciating mathematical and computational modeling, and in recent years several papers appeared in the literature: they have begun to investigate the various aspects of the immune system response to cancer from a computational and mathematical perspective [19–21].

Particularly, in [22], the authors developed a mathematical model to describe the growth dynamics of an immunogenic tumor in the presence of an active immune response. They paid special attention on the interaction of cancer cells with cytotoxic T lymphocytes and professional antigen presenting cells in a relatively small, multicellular tumor, before the angiogenesis process.

During the numerical simulation of the model, it has been discovered that adoptive immunotherapy protocols have the potential to promote tumor growth instead of inhibiting it. Conversely, active vaccination with tumor-antigen pulsed APCs was shown to be generally more effective.

7. Epitope Focused Immunoinformatics

DNA vaccination has been widely explored to develop new, alternative, and efficient vaccines for cancer immunotherapy. They offer several paybacks such as specific targeting, use of multiple genes to enhance immunity, and reduced risk compared to conventional vaccines.

Fast advances in molecular biology and immunoinformatics allow logical design methodologies. These technologies allow construction of DNA vaccines encoding selected tumor antigens together with molecules to direct and amplify the desired effector pathways, as well as highly targeted vaccines aimed at specific epitopes. Reliable predictions of immunogenic T cell epitope peptides are crucial for rational vaccine design and represent a key problem in immunoinformatics [23, 24].

For example, the authors in [25] explore the selection of T cell epitopes to develop epitope-based vaccines, the need for CD4+ T cell help for improved vaccines and the assessment of vaccine performance against tumor.

Moreover they present two applications, namely prediction of novel T cell epitopes and epitope enhancement by sequence modification, and combined rationale design with bioinformatics for creation of new synthetic mini-genes.

8. Repositories in Machine Learning Algorithms

It is well known that the immune system is characterized by high combinatorial complexity, especially due to its wide potential repertoire. Consequently, the analysis of immunological data needs the use of specialized computational tools.

A new way to select vaccine targets and reduce the number of necessary experiments is based on the use of machine learning (ML) algorithms in combination with classical experimentation. As the development of ML algorithms requires standardized data sets that are measured in a consistent way (and share the same uniform scale), there is a gap between the immunology community and the ML community. To overcome this problem and filling the gap, the authors in [26] present a repository for machine learning in immunology named Dana-Farber Repository for Machine Learning in Immunology (DFRMLI). Integrating experimental and in silico methods allows efficient study of highly combinatorial problems related with interpreting immune responses. With the advancement of experimental technologies, the amount of immunological data produced and distributed is increasing dramatically. Bioinformatics tools also based on statistical and ML algorithms are able to utilize these data. The main problem with both immunological data and other biological data is that they are usually represented qualitatively and as a consequence, these descriptions are often ambiguous, presenting a challenge for the mainstream ML developers. The DFRMLI is designed to overwhelm this hole through extending immunological data with well-defined annotations that could be conveniently used by the ML community.

9. Models in Flow Cytometry Data for Cancer Vaccine Immune Monitoring

Detection of minimal residual disease, diagnosis, characterization of the profile of immunotherapies, and immune response tracking represent hot topics in cancer research. Flow cytometry (FCM) is widely used in these areas of interest. Circumventing spurious positive events and recognizing uncommon cells subsets delineate the challenge in all these applications. To accomplish this is task, the use of multiple markers simultaneously in the analysis of FCM data will help a lot. This because the additional information provided often lends a hand to minimize the number of false positive and false negative events, hence improving both sensitivity and specificity.

With the use of the above explained strategy by manual gating, it is possible to analyze at most two markers in a single dot plot, often applying a sequential scheme. The sequential strategy is difficult to assess, as it gets rid of events that fall outside preceding gates at each stage.

Model-based analysis is a promising computational technique that works using information from all marker dimensions simultaneously and offers an alternative approach to flow analysis that can usefully complement manual gating in the design of optimal gating strategies. In [27], the authors presented results from model-based analysis illustrated with examples from FCM assays commonly used in cancer immunotherapy laboratories.

The authors' approach to model-based analysis is based on the use of statistical mixture models. Statistical mixture models are very widely used in the presence of problems where objects depicted in several or many dimensions need to be clustered or classified.

10. Modeling Personalized Response to Cancer Immunotherapy

Therapeutic interventions that stimulate tumor-specific immunity still remain rare. An improved understanding of patient-specific dynamic interactions of immunity and tumor progression, combined with personalized application of immune therapeutics, would increase the efficacy of immunotherapy. In [28] the authors developed a method to predict and enhance the individual response to immunotherapy by using personalized mathematical models. The approach is set in the early phase of treatment and includes an iterative real-time in-treatment evaluation of patient-specific parameters from the accruing clinical data, construction of personalized models and their validation, model-based simulation of subsequent response to ongoing therapy, and suggestion of potentially more effective patient-specific modified treatment. The model is then applied to a prostate cancer immunotherapy. The major finding of the simulations conducted in [28] suggested that an increase in vaccine dose and administration frequency would stabilize the disease in most patients.

11. Immunotherapies Enhancing Vaccines

Recently, Wilson and Levy [29] have investigated the possible effect of an immunotherapy based on an immunoregulatory protein, the transforming growth factor beta, (TGF-β), in combination with vaccine treatments. The proposed mathematical model follows the dynamics of the tumor size, TGF-β concentration, activated cytotoxic effector cells, and regulatory T cells. Using numerical simulations and stability analysis, they have studied several scenarios: a control case of no treatment, anti-TGF-β treatment, vaccine treatment, and combined anti-TGF-β vaccine treatments. The model was able to capture experimental results, and hence has the potential to be used in designing future experiments involving this approach to immunotherapy.

12. Conclusions

The investigation of vaccines and therapeutic approaches against cancer from the mathematical and computational point of view is still a new field of research. It has been shown that several papers have begun to propose models that have been appreciated by both clinicians and experimentalists and have been proven to be of great use in improving anti-cancer approaches research.

We expect that an extensive use of mathematical/computational modeling into clinical practice will stimulate the clinical research of new and alternative protocols for cancer treatments with immune interventions.

The possibility of the use of personalized approaches into the clinical practice is probably still far to come. However, virtual patient simulations can produce expected responses to the therapy for different class of patients (by immunological profile, age, pathologies, etc.). This can help the clinicians in deciding the best clinical approach for the specific patient.

Finally, a thought on the future directions of the modeling cancer vaccines topic. We believe that models should be integrated during the entire cycle of cancer vaccine development line. This means that if a model has been used in the first line of the development (for example in the definition of epitopes targets), it should be used later in the optimization of the schedule and finally in the human response to the specific vaccine or immunotherapy. Presently, to the best of our knowledge, there is no model that has been applied to the three critical phases of vaccine development.

References

[1] W. Zou, "Immunosuppressive networks in the tumour environment and their therapeutic relevance," *Nature Reviews Cancer*, vol. 5, no. 4, pp. 263–274, 2005.

[2] E. Gilboa, "The promise of cancer vaccines," *Nature Reviews Cancer*, vol. 4, no. 5, pp. 401–411, 2004.

[3] B. A. Guinn, N. Kasahara, F. Farzaneh, N. A. Habib, J. S. Norris, and A. B. Deisseroth, "Recent advances and current challenges in tumor immunology and immunotherapy," *Molecular Therapy*, vol. 15, no. 6, pp. 1065–1071, 2007.

[4] G. P. Dunn, L. J. Old, and R. D. Schreiber, "The three Es of cancer immunoediting," *Annual Review of Immunology*, vol. 22, pp. 329–360, 2004.

[5] J. Begley and A. Ribas, "Targeted therapies to improve tumor immunotherapy," *Clinical Cancer Research*, vol. 14, no. 14, pp. 4385–4391, 2008.

[6] M. Bhattacharya-Chatterjee, S. K. Chatterjee, and K. A. Foon, "Anti-idiotype antibody vaccine therapy for cancer," *Expert Opinion on Biological Therapy*, vol. 2, no. 8, pp. 869–881, 2002.

[7] F. O. Nestle, A. Farkas, and C. Conrad, "Dendritic-cell-based therapeutic vaccination against cancer," *Current Opinion in Immunology*, vol. 17, no. 2, pp. 163–169, 2005.

[8] P.-L. Lollini, F. Cavallo, P. Nanni, and G. Forni, "Vaccines for tumour prevention," *Nature Reviews Cancer*, vol. 6, no. 3, pp. 204–216, 2006.

[9] N. Kumar, B. S. Hendriks, K. A. Janes, D. de Graaf, and D. A. Lauffenburger, "Applying computational modeling to drug discovery and development," *Drug Discovery Today*, vol. 11, no. 17-18, pp. 806–811, 2006.

[10] Y. He, R. Rappuoli, A. S. De Groot, and R. T. Chen, "Emerging vaccine informatics," *Journal of Biomedicine and Biotechnology*, vol. 2010, Article ID 218590, 26 pages, 2010.

[11] F. Pappalardo, P.-L. Lollini, F. Castiglione, and S. Motta, "Modeling and simulation of cancer immunoprevention vaccine," *Bioinformatics*, vol. 21, no. 12, pp. 2891–2897, 2005.

[12] P. Nanni, G. Nicoletti, C. De Giovanni et al., "Combined allogeneic tumor cell vaccination and systematic interleukin 12 prevents mammary carcinogenesis in HER-2/neu transgenic mice," *Journal of Experimental Medicine*, vol. 194, no. 9, pp. 1195–1205, 2001.

[13] A. Palladini, G. Nicoletti, F. Pappalardo et al., "In silico modeling and in vivo efficacy of cancer-preventive vaccinations," *Cancer Research*, vol. 70, no. 20, pp. 7755–7763, 2010.

[14] F. Pappalardo, M. Pennisi, F. Castiglione, and S. Motta, "Vaccine protocols optimization: in silico experiences," *Biotechnology Advances*, vol. 28, no. 1, pp. 82–93, 2010.

[15] P. Nanni, G. Nicoletti, A. Palladini et al., "Antimetastatic activity of a preventive cancer vaccine," *Cancer Research*, vol. 67, no. 22, pp. 11037–11044, 2007.

[16] M. Pennisi, F. Pappalardo, A. Palladini et al., "Modeling the competition between lung metastases and the immune system using agents," *BMC Bioinformatics*, vol. 11, supplement 7, article S13, 2010.

[17] F. Pappalardo, E. Mastriani, P.-L. Lollini, and S. Motta, "Genetic algorithm against cancer," in *Proceedings of the 6th international conference on Fuzzy Logic and Applications (WILF '06)*, vol. 3849 of *Lecture Notes in Computer Science*, pp. 223–228, 2006.

[18] M. Pennisi, R. Catanuto, F. Pappalardo, and S. Motta, "Optimal vaccination schedules using simulated annealing," *Bioinformatics*, vol. 24, no. 15, pp. 1740–1742, 2008.

[19] F. Pappalardo, A. Palladini, M. Pennisi, F. Castiglione, and S. Motta, "Mathematical and computational models in tumor immunology," *Mathematical Modelling of Natural Phenomena*, vol. 7, no. 3, pp. 186–203, 2012.

[20] D. Alemani, F. Pappalardo, M. Pennisi, S. Motta, and V. Brusic, "Combining cellular automata and lattice boltzmann method to model multiscale avascular tumor growth coupled with nutrient diffusion and immune competition," *Journal of Immunological Methods*, vol. 376, no. 1-2, pp. 55–68, 2012.

[21] F. Pappalardo, I. M. Forero, M. Pennisi, A. Palazon, I. Melero, and S. Motta, "Simb16: modeling induced immune system response against b16-melanoma," *PLoS ONE*, vol. 6, no. 10, 2011.

[22] B. Joshi, X. Wang, S. Banerjee, H. Tian, A. Matzavinos, and M. A. J. Chaplain, "On immunotherapies and cancer vaccination protocols: a mathematical modelling approach," *Journal of Theoretical Biology*, vol. 259, no. 4, pp. 820–827, 2009.

[23] S. Mishra and S. Sinha, "Immunoinformatics and modeling perspective of T cell epitope-based cancer immunotherapy: a holistic picture," *Journal of Biomolecular Structure and Dynamics*, vol. 27, no. 3, pp. 293–306, 2009.

[24] D. S. DeLuca and R. Blasczyk, "The immunoinformatics of cancer immunotherapy," *Tissue Antigens*, vol. 70, no. 4, pp. 265–271, 2007.

[25] S. Iurescia, D. Fioretti, V. M. Fazio, and M. Rinaldi, "Epitope-driven DNA vaccine design employing immunoinformatics against B-cell lymphoma: a biotech's challenge," *Biotechnology Advances*, vol. 30, pp. 372–383, 2012.

[26] G. L. Zhang, H. H. Lin, D. B. Keskin, E. L. Reinherz, and V. Brusic, "Dana-Farber repository for machine learning in immunology," *Journal of Immunological Methods*, vol. 374, pp. 18–25, 2011.

[27] J. Frelinger, J. Ottinger, C. Gouttefangeas, and C. Chan, "Modeling flow cytometry data for cancer vaccine immune monitoring," *Cancer Immunology, Immunotherapy*, vol. 59, no. 9, pp. 1435–1441, 2010.

[28] Y. Kogan, K. Halevi-Tobias, M. Elishmereni, S. Vuk-Pavlovic, and Z. Agur, "Reconsidering the paradigm of cancer immunotherapy by computationally aided real-time personalization," *Cancer Research*, vol. 72, no. 9, pp. 2218–2227, 2012.

[29] S. Wilson and D. Levy, "A mathematical model of the enhancement of tumor Vaccine efficacy by immunotherapy," *Bulletin of Mathematical Biology*, vol. 74, pp. 1485–1500, 2012.